Technion - Israel Institute of Technology
Computer Science Department

Sequential Consistency versus Linearizability

by

H. Attiya and J.L. Welch

Technical Report #674
October 1991

Sequential Consistency versus Linearizability[*]

Hagit Attiya[†]
Department of Computer Science
The Technion
Haifa 32000, Israel

Jennifer L. Welch[‡]
Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599-3175

October 14, 1991

[*] This paper combines and unifies results that appear in preliminary form in [6] and [5].

[†] Email: hagit@cs.technion.ac.il. Part of this work was performed while the author was visiting DEC Cambridge Research Laboratory and the Laboratory for Computer Science, MIT, supported by ONR contract N00014-85-K-0168, by NSF grants CCR-8611442 and CCR-8915206, and by DARPA contracts N00014-89-J-1988 and N00014-87-K-0825.

[‡] Email: welch@cs.unc.edu. The work of this author was supported in part by NSF grant CCR-9010730, an IBM Faculty Development Award, and an NSF Presidential Young Investigator Award.

Abstract

The power of two well-known consistency conditions for shared memory multiprocessors, sequential consistency and linearizability, is compared. The cost measure studied is the worst-case response time in distributed implementations of virtual shared memory supporting one of the two conditions. Three types of shared memory objects are considered: read/write objects, FIFO queues, and stacks. In all cases, the worst-case response time is very sensitive to the assumptions that are made about the timing information available to the system. Under the strong assumption that processes have perfectly synchronized clocks, it is shown that sequential consistency and linearizability are equally costly: we present upper bounds for linearizability and matching lower bounds for sequential consistency. If clocks are only approximately synchronized, then for all three object types it is shown that linearizability is more expensive than sequential consistency: we present upper bounds for sequential consistency and larger lower bounds for linearizability. The upper bounds are shown by presenting algorithms that use atomic broadcast in a modular fashion. The lower bound proofs for the approximate case use the technique of "shifting", first introduced for studying the clock synchronization problem.

1 Introduction

A fundamental problem in concurrent computing is how to provide programmers with a useful model of logically shared data, without sacrificing performance. The model must specify how the data can be accessed and what guarantees are provided about the results. Shared memory is an attractive paradigm for communication among computing entities because it is familiar from the uniprocessor case, it can be considered more high level than message passing, and many of the classical solutions for synchronization problems were developed for shared memory (e.g., mutual exclusion [15]).

This problem arises in many situations at different levels of abstraction. These situations include implementing a single shared variable out of weaker shared variables, cache coherence, building multiprocessors (with both physical and distributed shared memory), and high-level applications for loosely-coupled distributed systems such as distributed file systems and transaction systems.

To enhance performance (e.g., response time, availability, or fault-tolerance), many implementations employ multiple copies of the same logical piece of shared data (*caching*). Also, multiple user programs must be able to execute "concurrently," either with interleaved steps, or truly in parallel. More complications arise because at some level, each access to shared data has duration in time, from its start to its end; it is not instantaneous.

Thus, the illusion of atomic operations on single copies of objects must be supported by a *consistency mechanism*. The consistency mechanism guarantees that although operations may be executed concurrently on various copies and have some duration, they will *appear* to have executed atomically, in some sequential order that is consistent with the order seen at individual processes.[1] When this order must preserve the global (external) ordering of non-overlapping operations, this consistency guarantee is called ***linearizability*** ([22]);[2] otherwise, the guarantee is called ***sequential consistency*** ([24]). Obviously, linearizability implies sequential consistency.

Sequential consistency and linearizability are two well-known consistency conditions. As the definitions of these two conditions are similar, it is important to study the relationships between them. In this paper we present a quantitative comparison of the costs to implement sequential consistency and linearizability in a non-bused distributed system. Distributed implementations are of great interest because of their ability to scale up in size. The comparison is based on time complexity — the inherent response time of the best possible distributed implementation supporting each consistency condition. That is, we present upper and lower bounds on the worst-case response time for performing an operation on an object.

We consider several types of shared objects in this paper. Most previous research concentrated on read/write objects. However, since read/write objects do not provide an expressive and convenient abstraction for concurrent programming (cf. [21]), many multiprocessors now

[1] This condition is similar in flavor to the notion of *serializability* from database theory ([8, 32]); however, serializability applies to *transactions* which aggregate many operations.

[2] Also called *atomicity* ([21, 25, 31]) in the case of read/write objects.

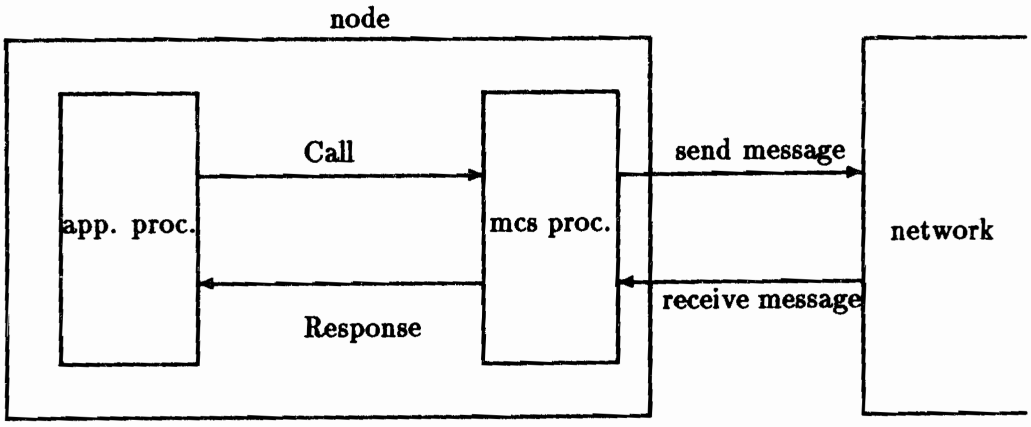


Figure 1: System Architecture

support more powerful concurrent objects, e.g., FIFO queues, stacks, test&set, fetch&add ([11]). Thus we also study FIFO queues and stacks.

We consider a collection of application programs running concurrently and communicating via virtual shared memory. The shared memory consists of a collection of *objects*. The application programs are running in a distributed system consisting of a collection of nodes and a complete communication network.[3] The shared memory abstraction is implemented by a *memory consistency system* (mcs), which uses local memory at the various nodes and some protocol executed by the mcs processes (one at each node). (Nodes that are dedicated storage can be modeled by nullifying the application process.) Fig. 1 illustrates a node, on which an application process and an mcs process are running. The application process sends calls to access shared data to the mcs process; the mcs process returns the responses to the application process, possibly based on messages exchanged with mcs processes on other nodes.

The correctness conditions are defined at the interface between the application processes (written by the user) and the mcs processes (supplied by the system). Thus, the mcs must provide the proper semantics when the values of the responses to calls are considered, throughout the network.

It turns out that the timing information available in the model has a crucial impact on the time complexity of implementing sequential consistency and linearizability. We assume that on each node there is a real-time clock readable by the mcs process at that node, that runs at the same rate as real-time. We assume that every message incurs a delay in the interval $[d - u, d]$, for some known constants u and d, $0 \leq u \leq d$ (u stands for *uncertainty*). If $u = 0$, then the message delays are constant.

First we consider the case when processes' clocks are perfectly synchronized. In this case, sequential consistency and linearizability are equally costly to implement, for the types of

[3] The assumption of a complete communication network can be omitted and is made here only for clarity of presentation.

objects we consider and for our cost measure. (Thus, separating sequential consistency from linearizability is not as obvious as it may seem.)

For read/write objects, we formalize and strengthen a result of Lipton and Sandberg [27], that in any sequentially consistent implementation, the sum of the worst-case response times for a read operation and a write operation is at least d. Obviously, this lower bound also holds for linearizable implementations. We then show that this tradeoff is tight—it is possible to have the response time of *only one* of the operations depend on the network's latency. Specifically, we present an algorithm in which a read operation is performed instantaneously (locally), while a write operation returns within time d; we also present an algorithm in which the roles are reversed. These algorithms achieve linearizability, and hence, sequential consistency.

For FIFO queues, we show that the worst-case response time for a dequeue operation is at least d. The result is proved for sequential consistency, and thus, holds also for linearizability. We show that this bound is tight by presenting an algorithm in which an enqueue operation returns instantaneously, while a dequeue operation returns within time d. The algorithm achieves linearizability, and hence, sequential consistency.

The situation for stacks is analogous to that for FIFO queues, with "pop" playing the role of "dequeue" and "push" the role of "enqueue".

We then turn to the more realistic case of approximately synchronized clocks. Under this assumption, for all three object types, there are gaps between the upper bounds for sequentially consistent implementations and the lower bounds for linearizable implementations. Operations that could be done instantaneously in the previous model can still be done instantaneously in sequentially consistent implementations, but they require $\Omega(u)$ time in linearizable implementations (note that u can be as large as d). Thus, under these timing assumptions linearizability is more expensive to implement than sequential consistency, when there are significantly more operations of one type.

For read/write objects, the lower bounds for linearizability are a worst-case time of $u/4$ for a read and a worst-case time of $u/2$ for a write. One sequentially consistent implementation guarantees time 0 for a read and time $2d$ for a write; another guarantees the reverse.

For FIFO queues, the lower bound for linearizability is a worst-case time of $u/2$ for an enqueue. Our sequentially consistent implementation guarantees time 0 for an enqueue and time $2d$ for a dequeue

As in the case of perfect clocks, the results for stacks are analogous to those for FIFO queues.

Our proofs make use of techniques from the theory of distributed systems. The lower bounds for implementations of linearizable objects are proved using *shifting* arguments, originally used in [28] for clock synchronization problems. Our efficient implementations of sequential consistency use as a subroutine a fast *atomic broadcast* algorithm ([10]) we have devised; however since our implementations are modular, any atomic broadcast algorithm will work.

Several papers have proposed sequentially consistent implementations of read/write objects, which were claimed to achieve a higher degree of concurrency (e.g., [2, 3, 7, 11, 16, 30, 34]).

In particular, Afek, Brown, and Merritt ([3]) present a sequentially consistent implementation of read/write objects, for systems where processes communicate via a bus. A bus enforces global ordering on all messages delivered to the processes; such a property is not provided in a communication network. None of these papers provides an analysis of the response time of the implementations suggested (or any other complexity measure). Furthermore, none of these papers proves that similar improvements cannot be achieved for linearizability. To the best of our knowledge, this is the first time such a result is shown and the first time other object types are considered.

This paper addresses a simplification of the problem of memory coherence in loosely-coupled multiprocessors ([7, 11, 9, 16, 26, 30, 33, 34]). Our formal model ignores several important practical issues, e.g., limitations on the size of local memory storage, network topology, clock drift and "hot-spots". Since our lower bounds are proved in a very strong model, they clearly hold for more practical systems. We believe our algorithms can be adapted to work in more realistic systems.

Section 2 presents our definitions and reviews the shifting technique. Section 3 considers the case of perfect clocks. There is one subsection for each of the three object types; each subsection consists of the lower bound(s) for sequential consistency followed by the upper bound(s) for linearizability. Section 4 covers the imperfect clock case. Again there is a subsection for each object type; now each subsection consists of the lower bound(s) for linearizability followed by the upper bound(s) for sequential consistency. We conclude in Section 5.

2 Preliminaries

2.1 Objects

Every shared object is assumed to have a ***serial specification*** (cf. [22]) defining a set of ***operations***, which are ordered pairs of call and response events, and a set of operation sequences, which are the allowable sequences of operations on that object. A sequence τ of operations for a collection of objects is ***legal*** if, for each object O, the restriction of τ to operations of O, denoted $\tau|O$, is in the serial specification of O.

In the case of a read/write object X, the ordered pair of events [$\text{Read}_p(X)$, $\text{Ret}_p(X, v)$] forms a Read operation for any process p and value v, and [$\text{Write}_p(X, v)$, $\text{Ack}_p(X)$] forms a Write operation. The set of operation sequences consists of all sequences in which every read operation returns the value of the latest preceding write operation (the usual read/write semantics).[4]

In the case of a FIFO queue Q, the ordered pair of events [$\text{Deq}_p(Q)$, $\text{Ret}_p(Q, v)$] forms a Deq operation for any process p and value v, and [$\text{Enq}_p(Q, v)$, $\text{Ack}_p(Q)$] forms an Enq

[4] The specifications used in this paper are operational. It is possible to give algebraic (axiomatic) specifications (cf. [22]); operational specifications are used here for simplicity.

operation. The set of operation sequences consists of all sequences that obey the usual FIFO queue semantics. That is, with a sequence of operations we associate a sequence of *queue states*, starting with an initial empty state and continuing with a state for each operation (representing the state of the queue *after* the operation. We require that each enqueue operation add an item to the end of the queue, and each dequeue operation remove an item from the head of the queue, or return $\perp$ if the queue is empty.

The specification of a stack S is similar to the specification of a queue: $[\text{Pop}_p(S), \text{Ret}_p(S, v)]$ forms a Pop operation for any process p and value v, and $[\text{Push}_p(S, v), \text{Ack}_p(S)]$ forms a Push operation. The set of operation sequences consists of all sequences that obey the usual (last-in-first-out) stack semantics.

2.2 System Model

We assume a system consisting of a collection of nodes connected via a communication network. On each node there is an application program, a memory-consistency system (mcs) process, and a real-time clock readable by the mcs process at that node. Formally, a *clock* is a monotonically increasing function from $\Re$ (real time) to $\Re$ (clock time).[5] The clock cannot be modified by the process. Processes do not have access to the real time; each process obtains its only information about time from its clock.

Below we list and informally explain the *events* that can occur at the mcs process on node p. (The name p is also used for the mcs process on node p).

1. *Call events:* the application program on node p wants to access a shared object.
2. *Response events:* the mcs process on node p is providing a response from a shared object to the application program on node p.
3. *Message receive events:* receive(p, m, q) for all messages m and nodes q: the mcs process on node p receives message m from the mcs process on node q.
4. *Message send events:* send(p, m, q) for all messages m and mcs processes q: the mcs process on node p sends message m to the mcs process on node q.
5. *Timer set events:* timerset(p, T) for all clock times T: p sets a timer to go off when its clock reads T.
6. *Timer events:* timer(p, T) for all clock times T: a timer that was set for time T on p's clock goes off.

The call, message-receive, and timer events are *interrupt events.*

[5] $\Re$ denotes the real numbers.

An *mcs process* (or simply *process*) is an automaton with a (possibly infinite) set of states, including an initial state, and a transition function. Each interrupt event causes an application of the transition function. The transition function is a function from states, clock times, and interrupt events to states, sets of response events, sets of message-send events, and sets of timer-set events (for subsequent clock times). That is, the transition function takes as input the current state, clock time, and interrupt event (which is the receipt of a call from the application process, or the receipt of a message from another node, or a timer going off), and produces a new state, a set of response events for the application process, a set of messages to be sent, and a set of timers to be set for the future.

A *step* of p is a tuple (s,T,i,s',R,M,S), where s and s' are states, T is a clock time, i is an interrupt event, R is a set of response events, M is a set of message-send events, S is a set of timer-set events, and s', R, M, and S are the result of p's transition function acting on s, T, and i.

A *history* of a process p with clock C is a mapping from $\Re$ (real time) to finite sequences of steps such that

1. for each real time t, there is only a finite number of times $t' < t$ such that the corresponding sequence of steps is nonempty (thus the concatenation of all the sequences in real-time order is a sequence);
2. the old state in the first step is p's initial state;
3. the old state of each subsequent step is the new state of the previous step;
4. for each real time t, the clock time component of every step in the corresponding sequence is equal to $C(t)$; and
5. for each real time t, in the corresponding sequence all non-timer events are ordered before any timer event and there is at most one timer event.

A *memory-consistency system (mcs)* is a set of processes P together with a set of clocks $\mathcal{C}$, one for each p in P. An *execution* of an mcs is a set of histories, one for each process p in P with clock C_p in $\mathcal{C}$, satisfying the following two conditions: (1) There is a one-to-one correspondence between the messages sent by p to q and the messages received by q from p, for any processes p and q. We use the message correspondence to define the *delay* of any message in an execution to be the real time of receipt minus the real time of sending. (2) A timer is received by p at clock time T if and only if p has previously set a timer for T. (The network is not explicitly modeled, although the constraints on executions imply that the network reliably delivers all messages sent.)

Execution σ is *admissible* if the following conditions hold:

1. For every p and q, every message in σ from p to q has its delay in the range $[d-u,d]$, for fixed nonnegative integers d and u, $u \leq d$. (This is a restriction on the network.)

2. For every p, at most one call at p is pending at a time. (This is a restriction on the application program.)

Note that the last condition allows each application program to have at most one call outstanding at any time. This outlaws pipelining or prefetching.

2.3 Correctness Conditions

Given an execution σ, let $ops(\sigma)$ be the sequence of call and response events appearing in σ in real-time order, breaking ties for each real time t as follows. First order all response events for time t whose matching call events occur before time t, using process ids to break any remaining ties. Then order all operations whose call and response both occur at time t. Preserve the relative ordering of operations for each process and break any remaining ties with process ids. Finally, order all call events for time t whose matching response events occur after time t, using process ids to break any remaining ties.

Our formal definitions of sequential consistency and linearizability follow. These definitions imply that every call gets an eventual response and that calls and responses alternate at each process. Given a sequence s of operations and a process p, we denote by $s|p$ the restriction of s to operations of p.

Definition 2.1 (Sequential consistency) *An execution σ is* sequentially consistent *if there exists a legal sequence τ of operations such that τ is a permutation of $ops(\sigma)$ and, for each process p, $ops(\sigma)|p$ is equal to $\tau|p$.*

Definition 2.2 (Linearizability) *An execution σ is* linearizable *if there exists a legal sequence τ of operations such that τ is a permutation of $ops(\sigma)$, for each process p, $ops(\sigma)|p$ is equal to $\tau|p$, and furthermore, whenever the response for operation op_1 precedes the call for operation op_2 in $ops(\sigma)$, then op_1 precedes op_2 in τ.*

An mcs is a *sequentially consistent* implementation of a set of objects if any admissible execution of the mcs is sequentially consistent; similarly, an mcs is a *linearizable* implementation of a set of objects if any admissible execution of the mcs is linearizable.

We measure the efficiency of an implementation by the worst-case response time for any operation on any object in the set. Given a particular mcs, an object O implemented by it, and an operation P on O, we denote by $|P(O)|$ the maximum time taken by a P operation on O in any admissible execution. We denote by $|P|$ the maximum of $|P(O)|$ over all objects O implemented by the mcs.

2.4 Shifting

A basic technique we use in our lower bound proofs (in Sections 4.1.1 and 4.2.1) is *shifting*, originally introduced in [28] to prove lower bounds on the precision achieved by clock synchronization algorithms. Shifting is used to change the timing and the ordering of events in the system while preserving the local views of the processes.

Informally, given an execution with a certain set of clocks, if process p's history is changed so that the real times at which the events occur are shifted by some amount s and if p's clock is shifted by the same amount, then the result is another execution in which every process still "sees" the same events happening at the same real time. The intuition is that the changes in the real times at which events happen at p cannot be detected by p because its clock has changed by a corresponding amount.

More precisely, the *view* of process p in history π of p with clock C is the concatenation of the sequences of steps in π, in real-time order. The real times of occurrence are not represented in the view. Two histories, one of process p with clock C and the other of process p with clock C', are *equivalent* if the view of p is the same in both histories. Two executions, execution σ of system (P, C) and execution σ' of (P, C'), are *equivalent* if for each process p, the component histories for p in σ and σ' are equivalent. Thus, the executions are indistinguishable to the processes. Only an outside observer who has access to the real time can tell them apart.

Given history π of process p with clock C, and real number s, a new history $\pi' = \mathit{shift}(\pi, s)$ is defined by $\pi'(t) = \pi(t+s)$ for all t. That is, all tuples are shifted earlier in π' by s if s is positive, and later by $-s$ if s is negative. Given a clock C and real number s, a new clock $C' = \mathit{shift}(C, s)$ is defined by $C'(t) = C(t) + s$ for all t. That is, the clock is shifted forward by s if s is positive, and backward by $-s$ if s is negative.

The following lemma observes that shifting a history of process p and p's clock by the same amount produces another history.

Lemma 2.1 *Let π be a history of process p with clock C, and let s be a real number. Then $\mathit{shift}(\pi, s)$ is a history of p with clock $\mathit{shift}(C, s)$.*

Given execution σ of system (P, C), and real number s, a new execution $\sigma' = \mathit{shift}(\sigma, p, s)$ is defined by replacing π, p's history in σ, by $\mathit{shift}(\pi, s)$, and by retaining the same correspondence between sends and receives of messages. (Technically, the correspondence is redefined so that a pairing in σ that involves the event for p at time t, in σ' involves the event for p at time $t - s$.) All tuples for process p are shifted by s, but no others are altered. Given a set of clocks $C = \{C_q\}_{q \in P}$, and real number s, a new set of clocks $C' = \mathit{shift}(C, p, s)$, is defined by replacing clock C_p by clock $\mathit{shift}(C_p, s)$. Process p's clock is shifted forward by s, but no other clocks are altered.

The following lemma observes that shifting one process' history and clock by the same amount in an execution results in another execution that is equivalent to the original.

Lemma 2.2 (Lundelius and Lynch) *Let σ be an execution of system (P,C), p a process, and s a real number. Let $C' = shift(C,p,s)$ and $\sigma' = shift(\sigma,p,s)$. Then σ' is an execution of (P,C'), and σ' is equivalent to σ.*

The following lemma quantifies how message delays change when an execution is shifted. Notice that the result of shifting an admissible execution is not necessarily admissible.

Lemma 2.3 (Lundelius and Lynch) *Let σ be an execution of system (P,C), p a process, and s a real number. Let $C' = shift(C,p,s)$ and $\sigma' = shift(\sigma,p,s)$. Make the obvious correspondence between messages in σ and in σ'. Suppose x is the delay of message m from process q to process r in σ. Then the delay of m in σ' is x if $q \neq p$ and $r \neq p$, $x - s$ if $r = p$, and $x + s$ if $q = p$.*

3 Perfect Clocks

We start by considering the case in which processes have perfectly synchronized (*perfect*) clocks and message delay is constant and known.[6] Perfect clocks are modeled by letting $C_p(t) = t$ for all p and t. The constant message delay is modeled by letting $u = 0$; d is known and can be used by the mcs.

For each of the three object types, we first prove lower bounds on the worst-case response time for sequentially consistent implementations. Since sequential consistency is a weaker condition than linearizability, these bounds also hold for linearizable implementations. Then we present algorithms that achieve linearizability, and hence sequential consistency, with worst-case response times matching the lower bounds. Section 3.1 considers read/write objects, Section 3.2 considers FIFO queues, and Section 3.3 considers stacks.

3.1 Read/Write Objects

We show in Section 3.1.1 that for sequential consistency the sum of the worst-case response times of read and write operations is at least d, even in this strong model. This is a formalization of a result of Lipton and Sandberg ([27, Theorem 1]), making precise the timing assumptions made on the system. We then show in Section 3.1.2 that the lower bound is tight for this model by describing two linearizable algorithms that match the lower bound exactly: In the first algorithm, reads are performed instantaneously, while the worst-case response time for a

[6] The assumptions that processes have perfect clocks and that message delays are constant (and known) are equivalent. If one assumes that clocks are not necessarily synchronized perfectly (but run at the rate of real time) and that the message delay is constant and known, then a simple algorithm suffices to synchronize the clocks perfectly. If one assumes that clocks are perfectly synchronized and that there is a known upper bound d on message delays, then constant message delays can be easily simulated by timestamping each message with the clock time of the sender and having each recipient delay any message that arrives with delay smaller than d until the delay is exactly d.

write is d. In the second algorithm, writes are performed instantaneously, while the worst-case response time for a read is d.

3.1.1 Lower Bounds for Sequential Consistency

Theorem 3.1 (Lipton and Sandberg) *For any memory-consistency system that is a sequentially consistent implementation of two read/write objects X and Y, $|Write| + |Read| \geq d$.*

Proof: Let p and q be two processes that access X and Y. Assume by way of contradiction that there exists a sequentially consistent implementation of X and Y for which both $|Write(X)| + |Read(Y)| < d$ and $|Write(Y)| + |Read(X)| < d$. Without loss of generality, assume that 0 is the initial value of both X and Y.

By the specification of Y, there is some admissible execution α_1 such that $ops(\alpha_1)$ is

$$\text{Write}_p(X,1)\ \text{Ack}_p(X)\ \text{Read}_p(Y)\ \text{Ret}_p(Y,0)$$

and $\text{Write}_p(X,1)$ occurs at real time 0 and $\text{Read}_p(Y)$ occurs immediately after $\text{Ack}_p(X)$. By assumption, the real time at the end of α_1 is less than d. Thus no message is received at any node during α_1.

By the specification of X, there is some admissible execution α_2 such that $ops(\alpha_2)$ is

$$\text{Write}_q(Y,1)\ \text{Ack}_q(Y)\ \text{Read}_q(X)\ \text{Ret}_q(X,0)$$

and $\text{Write}_q(Y,1)$ occurs at real time 0 and $\text{Read}_q(X)$ occurs immediately after $\text{Ack}_q(Y)$. By assumption, the real time at the end of α_2 is less than d. Thus no message is received at any node during α_2.

Since no message is ever received in α_1 and α_2, the execution α obtained from α_1 by replacing q's history with q's history in α_2 is admissible. Then $ops(\alpha)$ consists of the operations [$\text{Write}_p(X,1)$, $\text{Ack}_p(X)$] followed by [$\text{Read}_p(Y)$, $\text{Ret}_p(Y,0)$], and [$\text{Write}_q(Y,1)$, $\text{Ack}_q(Y)$] followed by [$\text{Read}_q(X)$, $\text{Ret}_q(X,0)$].

By assumption, α is sequentially consistent. Thus there is a legal operation sequence τ consisting of the operations [$\text{Write}_p(X,1)$, $\text{Ack}_p(X)$] followed by [$\text{Read}_p(Y)$, $\text{Ret}_p(Y,0)$], and [$\text{Write}_q(Y,1)$, $\text{Ack}_q(Y)$] followed by [$\text{Read}_q(X)$, $\text{Ret}_q(X,0)$]. Since τ is a sequence of operations, either the read of X follows the write of X, or the read of Y follows the write of Y. But each possibility violates the serial specification of either X or Y, contradicting τ being legal. ■

3.1.2 Upper Bounds for Linearizability

In this section we show that the tradeoff indicated by Theorem 3.1 is inherent, and that a sequentially consistent implementation may choose which operation to slow down. More precisely, we present an algorithm in which a read operation is instantaneous (local) while a write operation returns within time d; we also present an algorithm in which the roles are reversed. These algorithms actually ensure the stronger condition of linearizability.

The algorithm for fast reads and slow writes works as follows. Each process keeps a copy of all objects in its local memory. When a $\text{Read}_p(X)$ occurs, p reads the value v of X in its local memory and immediately does a $\text{Ret}_p(X, v)$. When a $\text{Write}_p(X, v)$ occurs, p sends "write(X, v)" messages to all other processes. Then p waits d time, after which it changes the value of X to v in its local memory and does an $\text{Ack}_p(X)$. Whenever a process receives a "write(X, v)" message, it changes the value of X to v in its local memory. (If it receives several at the same time, it "breaks ties" using sender ids; that is, it writes the value in the message from the process with the largest id and ignores the rest of the messages.)

Theorem 3.2 *There exists a linearizable implementation of read/write objects with* $|Read| = 0$ *and* $|Write| = d$.

Proof: Consider the algorithm just described. Clearly the time for every read is 0 and the time for every write is d.

Let σ be an admissible execution of this algorithm. For each operation in σ, say that it *occurs* at the real time when its response happens. Let τ be the sequence of operations in σ ordered by time of occurrence, breaking ties with process ids. Clearly $\sigma|p$ is equal to $\tau|p$ for all p, and the order of non-overlapping operations is preserved.

It remains to show that τ is legal, i.e., that for every object X, $\tau|X$ is in the serial specification of X. Since X is a read/write object, we must show that every Read Returns the value written by the latest preceding Write (and if there is no such Write, then it returns the initial value).

Pick any X and consider $\tau|X = op_1 op_2 \ldots$ Suppose op_i is $[\text{Read}_p(X), \text{Ret}_p(X, v)]$ and op_i occurs at time t in σ.

Case 1: No Write precedes op_i in τ. By the definition of τ, no Write is Acked before op_i starts. Since the Ack for a Write happens at the same time that every process updates its local copy of X, the Read reads the initial value for X and Returns that value.

Case 2: Some $\text{Write}_p(X, v)$ is the latest Write preceding op_i in τ. By the definition of τ, this Write is Acked before op_i starts, but no other Write is Acked before op_i starts. Since the Ack for a Write happens at the same time that every process updates its local copy of X, the Read reads v for the value of X and Returns that value. ■

The algorithm for slow reads and fast writes is similar to the previous one. Each process keeps a copy of all objects in its local memory. When a Read$_p(X)$ occurs, p waits d time, after which it reads the value v of X in its local memory and immediately does a Ret$_p(X, v)$. When a Write$_p(X, v)$ occurs, p sends "write(X, v)" messages to all other processes (including a dummy message to itself which is delayed d time) and does an Ack immediately. Whenever a process receives a "write(X, v)" message, it changes the value of X to v in its local memory. Ties are resolved as in the previous algorithm.

Theorem 3.3 *There exists a linearizable implementation of read/write objects with $|Read| = d$ and $|Write| = 0$.*

Proof: Consider the algorithm just described. Clearly the time for every read is d and the time for every write is 0.

Let σ be an admissible execution of this algorithm. For each operation in σ, say that it *occurs* at the real time when its call happens. Let τ be the sequence of operations in σ ordered by time of occurrence, breaking ties with process ids. Clearly $\sigma|p$ is equal to $\tau|p$ for all p, and the order of non-overlapping operations is preserved.

It remains to show that τ is legal, i.e., that for every object X, $\tau|X$ is in the serial specification of X. Since X is a read-write object, we must show that every Read Returns the value written by the latest preceding Write (and if there is no such Write, then it returns the initial value).

Pick any X and consider $\tau|X = op_1 op_2 \ldots$. Suppose op_i is [Read$_p(X)$, Ret$_p(X, v)$] and op_i occurs at time t in σ.

Case 1: No Write precedes op_i in t. By the definition of τ, no Write starts before op_i starts. Since the local changes occur d time after the Write starts and the Read reads the local memory d time after the Read starts, it reads the local memory before any change is made to it. Thus the Read returns the initial value.

Case 2: Some Write$_p(X, v)$ is the latest Write preceding op_i in τ. Essentially the same argument as in Case 1 works. ■

3.2 FIFO Queues

We show in Section 3.2.1 that for sequential consistency the worst-case response time of a dequeue operation is at least d, even when clocks are perfectly synchronized and message delays are constant. We then show in Section 3.2.2 that this lower bound is tight for this model by describing a linearizable algorithm that matches the lower bound exactly: enqueues are performed instantaneously, while dequeues take time d.

3.2.1 Lower Bound for Sequential Consistency

Theorem 3.4 *For any sequentially consistent implementation of a FIFO queue Q, $|Deq(Q)| \geq d$.*

Proof: Let p and q be two processes that access Q. Assume by way of contradiction that there exists a sequentially consistent implementation of Q for which $|Deq(Q)| < d$. Let $T = |Deq(Q)|$. By definition, the queue Q is initially empty.[7] By the specification of Q, there is some admissible execution α'_1 such that $ops(\alpha'_1)$ is

$$\mathrm{Enq}_q(Q,1)\ \mathrm{Ack}_q(Q)\ \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,v_1)\ \ldots \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,v_i)\ \ldots$$

$\mathrm{Enq}_q(Q,1)$ occurs at real time 0 and $\mathrm{Ack}_p(Q)$ occurs at time t; the first $\mathrm{Deq}_p(Q)$ occurs at time t, while the jth $\mathrm{Deq}_p(Q)$ occurs at time $t+(j-1)T$ (see Figure 2(a)[8]). Consider now the infinite sequence $v_1,\ldots,v_i,\ldots$. It is possible that many of them are $\perp$; however, since only a finite number of Deq operations can be serialized before the Enq operation, we have:

Lemma 3.5 *There exists some i such that $v_i \neq \perp$.*

Fix this particular i, and note that $v_i = 1$ and, for all j, $1 \leq j < i$, $v_j = \perp$. Let α_1 be α'_1 truncated after the ith Deq operation by p. More precisely, $ops(\alpha_1)$ is

$$\mathrm{Enq}_q(Q,1)\ \mathrm{Ack}_p(Q)\ \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,\perp)\ \ldots \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,\perp)\ \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,1)$$

$\mathrm{Enq}_q(Q,1)$ occurs at real time 0 and $\mathrm{Ack}_p(Q)$ occurs at time t; the first $\mathrm{Deq}_p(Q)$ occurs at time t, while the ith $\mathrm{Deq}_p(Q)$ occurs at time $t+(i-1)T$ (see Figure 2(b)). It is clear that the v_j's are exactly as in α'_1. By assumption, the real time at the end of α is less than $t+(i-1)T+d$. Thus, no message sent after $t+(i-1)T$ is received during α_1.

We now consider the execution where the ith (and last) dequeue by p is replaced with a dequeue by q. More precisely, by the specification of Q, there is some admissible execution α_2 such that $ops(\alpha_2)$ is

$$\mathrm{Enq}_q(Q,1)\ \mathrm{Ack}_p(Q)\ \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,\perp)\ \ldots \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,\perp)\ \mathrm{Deq}_q(Q)\ \mathrm{Ret}_q(Q,u)$$

$\mathrm{Enq}_q(Q,1)$ occurs at real time 0 and $\mathrm{Ack}_p(Q)$ occurs at time t; the first $\mathrm{Deq}_p(Q)$ occurs at time t, while the $(i-1)$st $\mathrm{Deq}_p(Q)$ occurs at time $t+(i-2)T$, and Deq_q occurs at time $t+(i-1)T$ (see Figure 2(c)). Since α_2 is sequentially consistent, it follows that $u = 1$. By assumption, the

[7] If we allow queues to be initially non-empty, the proof of the lower bound becomes much simpler; we leave the details to the interested reader.

[8] In the figures, time runs from left to right, and each line represents events at one process. Important time points are marked at the bottom.

real time at the end of α_2 is less than $t+(i-1)T+d$. Thus, no message sent after $t+(i-1)T$ is received during α_2.

Since no message sent after time $t+(i-1)T$ is ever received in α_1 and α_2, and since α_1 and α_2 are identical until time $t+(i-1)T$, the execution α obtained from α_1 by replacing q's history with q's history in α_2 is admissible. Then $ops(\alpha)$ is

$$\mathrm{Enq}_q(Q,1)\ \mathrm{Ack}_p(Q)\ \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,\perp)\ \ldots \mathrm{Deq}_p(Q)\ \mathrm{Ret}_p(Q,1)\ \mathrm{Deq}_q(Q)\ \mathrm{Ret}_q(Q,1)$$

(see Figure 2(d)). By assumption, α is sequentially consistent. Thus, there is a legal sequence τ, which is a permutation of the above operations. However, in τ the element "1" is enqueued once but dequeued twice, a contradiction. ■

3.2.2 Upper Bound for Linearizability

In this section we show that the lower bound given in Theorem 3.4 is tight for the model with perfect clocks. Specifically, we present an algorithm in which an enqueue operation returns instantaneously, while a dequeue operation returns within time d. The algorithm ensures the stronger condition of linearizability.

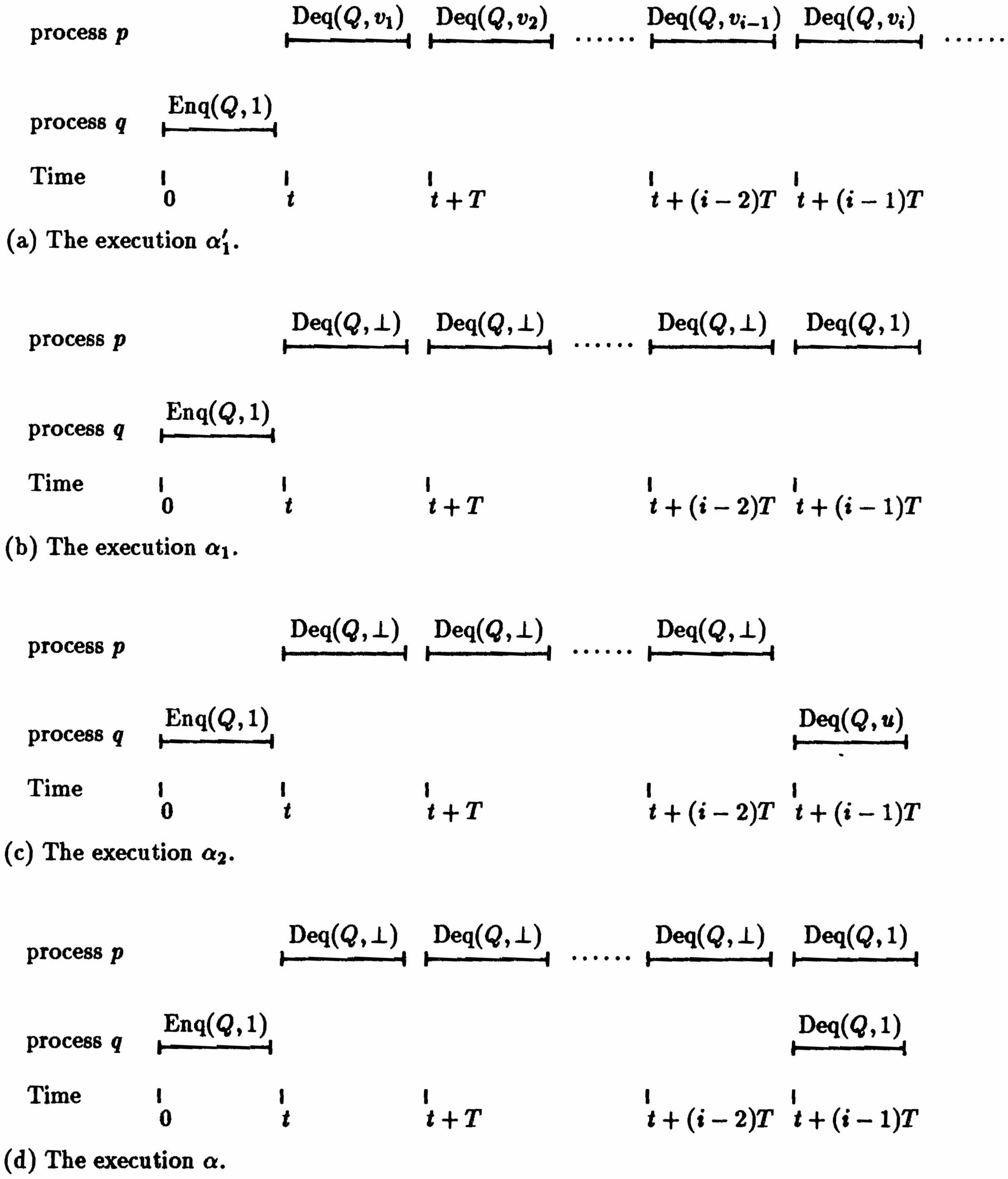
The algorithm works as follows. Each process keeps a copy of all queues in its local memory. When an $\mathrm{Enq}_p(Q,v)$ occurs, p sends "enqueue(Q,v)" messages to all other processes (including a message to itself which is delayed d time) and does an Ack immediately. When a $\mathrm{Deq}_p(Q)$ occurs, p sends "dequeue(Q)" messages to all other processes (including a message to itself which is delayed d time). After waiting d time, p handles its own message and does a $\mathrm{Ret}_p(Q,v)$. Whenever a process receives an "enqueue(Q,v)" or "dequeue(Q)" message, it makes the appropriate update to the copy of Q in its local memory. (If it receives several messages at the same time, it "breaks ties" using sender ids, that is, it handles them by increasing order of process ids.)

Theorem 3.6 *There exists a linearizable implementation of FIFO queues with $|Enq| = 0$ and $|Deq| = d$.*

In the proof, we serialize each operation to occur d time after it is called. Since all processes update their local copies at these serialization times, the claim follows.

Proof: Consider the algorithm just described. Clearly $|Enq| = 0$ and $|Deq| = d$.

Let σ be an admissible execution of this algorithm. For each operation in σ, say that it *occurs* at time d after the real time when its call happens. Let τ be the sequence of operations in σ ordered by time of occurrence, breaking ties with process ids. Clearly $\sigma|p$ is equal to $\tau|p$ for all p, and the order of non-overlapping operations is preserved.



(a) The execution α'_1.

(b) The execution α_1.

(c) The execution α_2.

(d) The execution α.

Figure 2: Executions used in the proof of Theorem 3.5.

It remains to show that τ is legal, i.e., that for every object Q, $\tau|Q$ is in the serial specification of Q. Pick any Q and consider $\tau|Q = op_1 op_2 \ldots$. Suppose op_i is [$\text{Deq}_p(Q)$, $\text{Ret}_p(Q, v)$]. Because message delay is fixed, updates at p to the local copy of Q occur in the same order as in τ, and the claim follows. ■

3.3 Stacks

The results for stacks are analogous to those for FIFO queues, with Pop playing the role of Deq and Push the role of Enq.

Theorem 3.7 *For any sequentially consistent implementation of a stack S, $|Pop(S)| \geq d$.*

Theorem 3.8 *There exists a linearizable implementation of stacks with $|Push| = 0$ and $|Pop| = d$.*

4 Imperfect Clocks

Obviously, the assumptions of the previous section are unrealistically strong. In this section we relax them, and assume a system in which clocks run at the same rate as real time but are not initially synchronized, and in which message delays are in the range $[d - u, d]$ for some $u > 0$.

Under these assumptions, the lower bounds of Section 3 still hold, but the algorithms of Section 3 do not work. We show that in this model there is a gap between the upper bounds for sequential consistency and the lower bounds for linearizability, for all three object types. The new lower bounds for linearizability show that operations that could be done instantaneously in the previous model now require at least $\Omega(u)$ time. Recall that u is the uncertainty in the message delay and can be as large as d. The new upper bounds are algorithms that match, within constant factors, the lower bounds for sequential consistency in the previous model.

Inspecting the algorithms from Section 3 reveals that in all cases correctness hinges on the fact that updates are handled by all processes in the same order and at the same time. In order to guarantee sequential consistency, it suffices for processes to update their local copies in the same order (not necessarily at the same time). A simple way to achieve this property is for a centralized controller to collect update messages and broadcast them. Using atomic broadcast it is possible to translate this idea into algorithms that are fully distributed and do not rely on a centralized controller. The algorithms are completely asynchronous, and do not rely on timing information.

Atomic broadcast ([10]) is a communication primitive which guarantees that every message sent using the primitive is received at every process, that all messages are delivered in the same order at all processes, and that two messages sent by the same process are delivered in the same order they were sent. Our implementations are described in a modular way so that they will work with any atomic broadcast algorithm (e.g., [10, 13, 19]). The interface to the primitive consists of two operations, ABC-send(m) to send a message m (possibly consisting of several fields) and ABC-receive(m) to receive a message m. In analyzing our implementations, we assume there is a known bound, h, on the time that the atomic broadcast primitive takes to deliver a message to all processes. Each of our implementations has one fast operation, which takes time 0, and one slow operation, which takes time h. In Appendix A we describe and prove correct a fast atomic broadcast algorithm with $h = 2d$. By using this algorithm in our impelmentations, we obtain implementations in which slow operations take time $2d = O(d)$.

4.1 Read/Write Objects

We show in Section 4.1.1 that in any linearizable implementation of a read/write object, the worst-case response time of *both* read and write operations must depend on u. We then present in Section 4.1.2 two algorithms for read/write objects, one in which reads are performed instantaneously while the worst-case response time for a write is $O(d)$, and another in which the roles are reversed.

4.1.1 Lower Bounds for Linearizability

We now show that, under reasonable assumptions about the pattern of sharing, in any linearizable implementation of an object, the worst-case time for a read is $u/4$ and the worst-case time for a write is $u/2$. The proofs of these lower bounds use the technique of shifting, described in Section 2.4.

Theorem 4.1 *Assume X is a read/write object with at least two readers. Then any linearizable implementation of X must have $|Read(X)| \geq \frac{u}{4}$.*

Proof: Let p and q be two processes that read X and r be a process that writes X. Assume in contradiction that there is an implementation with $|Read(X)| < \frac{u}{4}$. Without loss of generality, assume that the initial value of X is 0. The idea of the proof is to consider an execution in which p reads 0 from X, then q and p alternate reading X while r writes 1 to X, and then q reads 1 from X. Thus there exists a read R_1, say by p, that returns 0 and is immediately followed by a read R_2 by q that returns 1. If q is shifted earlier by $u/2$, then R_2 precedes R_1 in the resulting execution. Since R_2 returns the new value 1 and R_1 returns the old value 0, this contradicts linearizability.

Let $k = \lceil \frac{|Write(X)|}{u} \rceil$. By the specification of X, there is an admissible execution α, in which all message delays are $d - \frac{u}{2}$, consisting of the following operations (see Fig. 3(a)):

- At time $\frac{u}{4}$, r does a Write$_r(X, 1)$.

- Between times $\frac{u}{4}$ and $(4k+1)\cdot\frac{u}{4}$, r does an $\mathrm{Ack}_r(X)$. (By definition of k, $(4k+1)\cdot\frac{u}{4} \geq \frac{u}{4} + |Write(X)|$, and thus r's write operation is guaranteed to finish in this interval.)
- At time $2i\cdot\frac{u}{4}$, p does a $\mathrm{Read}_p(X)$, $0 \leq i \leq 2k$.
- Between times $2i\cdot\frac{u}{4}$ and $(2i+1)\cdot\frac{u}{4}$, p does a $\mathrm{Ret}_p(X, v_{2i})$, $0 \leq i \leq 2k$.
- At time $(2i+1)\cdot\frac{u}{4}$, q does a $\mathrm{Read}_q(X)$, $0 \leq i \leq 2k$.
- Between times $(2i+1)\cdot\frac{u}{4}$ and $(2i+2)\cdot\frac{u}{4}$, q does a $\mathrm{Ret}_q(X, v_{2i+1})$, $0 \leq i \leq 2k$.

Thus in $ops(\alpha)$, p's read of v_0 precedes r's write, q's read of v_{4k+1} follows r's write, no two read operations overlap, and the order of the values read from X is v_0, v_1, v_2, ..., v_{4k+1}. By linearizability, $v_0 = 0$ and $v_{4k+1} = 1$. Thus there exists j, $0 \leq j \leq 4k$, such that $v_j = 0$ and $v_{j+1} = 1$. Without loss of generality, assume that j is even, so that v_j is the result of a read by p.

Define $\beta = \mathrm{shift}(\alpha, q, \frac{u}{2})$; i.e., we shift q earlier by $\frac{u}{2}$. (See Fig. 3(b).) The result is admissible, since by Lemma 2.3 the message delays to q become $d-u$, the message delays from q become d, and the remaining message delays are unchanged.

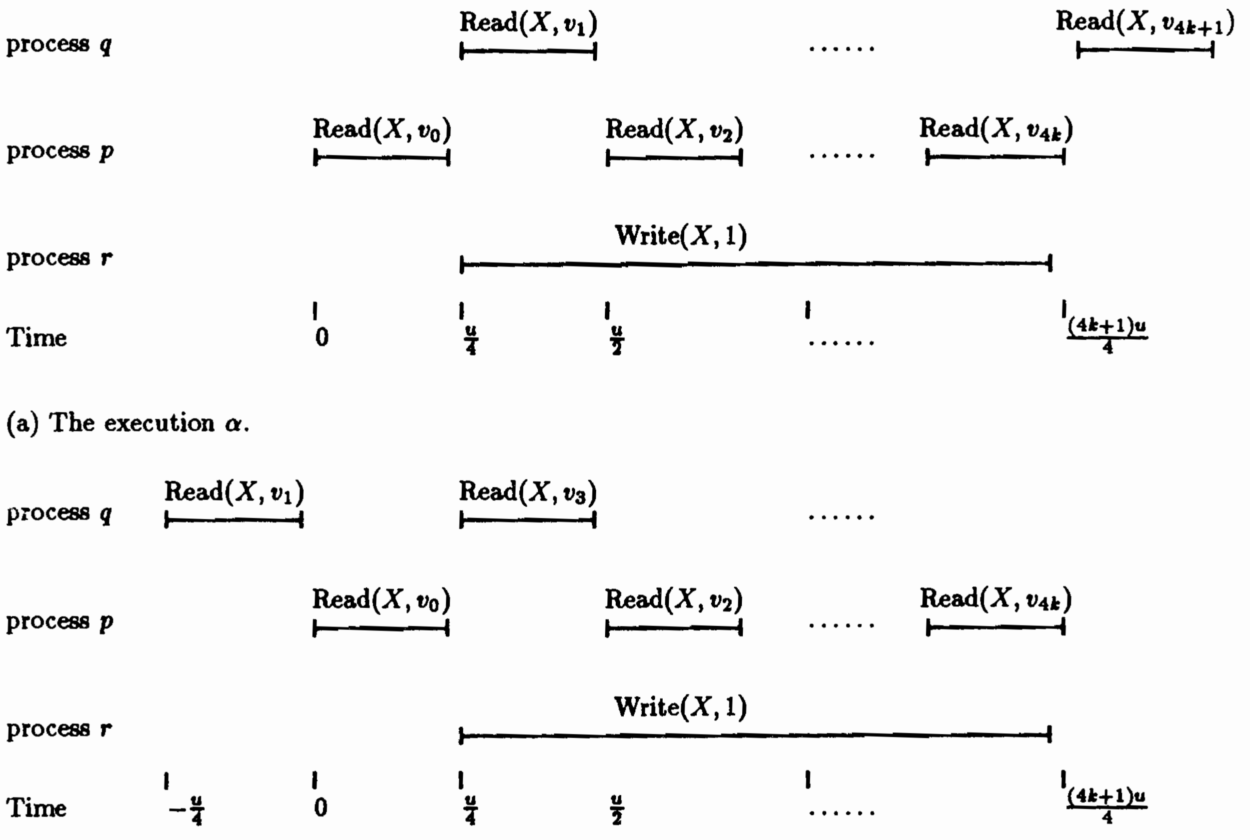
As a result of the shifting, we have reordered read operations with respect to each other at p and q. Specifically, in $ops(\beta)$, the order of the values read from X is v_1, v_0, v_3, v_2, ..., v_{j+1}, v_j, ... Thus in β we now have $v_{j+1} = 1$ being read before $v_j = 0$, which violates linearizability. ■

Theorem 4.2 *If X is a read/write object with at least two writers, then any linearizable implementation of X must have $|Write(X)| \geq \frac{u}{2}$.*

The proof uses techniques similar to the proof of Theorem 4.1. It constructs an execution in which, if write operations are too short, linearizability can be violated by appropriately shifting histories.

Proof: Let p and q be two processes that write X and r be a process that reads X. Assume in contradiction that there is an implementation with $|Write(X)| < \frac{u}{2}$. Without loss of generality, assume that the initial value of X is 0. By the specification of X, there is an admissible execution α such that

- $ops(\alpha)$ is $\mathrm{Write}_p(X,1)$ $\mathrm{Ack}_p(X)$ $\mathrm{Write}_q(X,2)$ $\mathrm{Ack}_q(X)$ $\mathrm{Read}_r(X)$ $\mathrm{Ret}_r(X,2)$;
- $\mathrm{Write}_p(X,1)$ occurs at time 0, $\mathrm{Write}_q(X,2)$ occurs at time $\frac{u}{2}$, and $\mathrm{Read}_r(X)$ occurs at time u; and
- the message delays in α are d from p to q, $d-u$ from q to p, and $d-\frac{u}{2}$ for all other ordered pairs of processes.



(a) The execution α.

(b) The execution β.

Figure 3: Executions used in the proof of Theorem 4.1.

Let $\beta = \mathit{shift}(\mathit{shift}(\alpha, p, -\frac{u}{2}), q, \frac{u}{2})$; i.e., we shift p later by $\frac{u}{2}$ and q earlier by $\frac{u}{2}$. The result is still an admissible execution, since by Lemma 2.3 the delay of a message from p or to q becomes $d - u$, the delay of a message from q or to p becomes d, and all other delays are unchanged.

But $\mathit{ops}(\beta)$ is

$$\mathrm{Write}_q(X,2)\ \mathrm{Ack}_q(X)\ \mathrm{Write}_p(X,1)\ \mathrm{Ack}_p(X)\ \mathrm{Read}_r(X)\ \mathrm{Ret}_r(X,2)$$

which violates linearizability, because r's read should return 1, not 2. ■

The assumptions about the number of readers and writers made in Theorems 4.1 and 4.2 are crucial to the results, since it can be shown that the algorithms from Theorems 3.2 and 3.3 are correct if there is only one reader and one writer.

```
Read_p(X):
  generate Ret_p(X, v), where v is the value of p's copy of X

Write_p(X, v):
  ABC-send(X, v)

ABC-receive(X, v) from q:
  set local copy of X to v
  if q = p then generate Ack_p(X) endif
```

Figure 4: Sequentially consistent fast read algorithm.

4.1.2 Upper Bounds for Sequential Consistency

Fast Reads

We start with the algorithm for fast reads (time 0) and slow writes (times at most h).

In the algorithm, each process keeps a local copy of every object. A read returns the value of the local copy immediately. When a write comes in to p, p sends an atomic broadcast containing the name of the object to be updated and the value to be written; but it does not yet generate an Ack for the write operation. When an update message is delivered to a process q, q writes the new value to its local copy of the object. If the update message was originated by q, then q generates an Ack and the (unique pending) write operation returns.

More precisely, the state of each process consists of a copy of every object, initially equal to its initial value. The transition function of process p appears in Fig. 4.

To prove the correctness of the algorithm, we first show:

Lemma 4.3 *For every admissible execution and every process p, p's local copies take on all the values contained in write operations, all updates occur in the same order at each process, and this order preserves the order of write operations on a per-process basis.*

Proof: By the code, an ABC-send is done exactly once for each write operation. By the guarantees of the atomic broadcast, each process receives exactly one message for each write operation, these messages are received in the same order at each process, and this order respects the order of sending on a per-process basis. ■

Call the total order of Lemma 4.3 the "Abcast order".

Lemma 4.4 *For every admissible execution, every process p, and all objects X and Y, if read R of object Y follows write W to object X in $ops(\sigma)|p$, then R's read of p's local copy of Y follows W's write of p's local copy of X.*

Proof: The lemma is true because W does not end until its update is performed at its initiator. ∎

Theorem 4.5 *There exists a sequentially consistent implementation of read/write objects with* $|Read| = 0$ *and* $|Write| = h$.

Proof: Consider the algorithm just presented. Clearly the time for any read is 0. The time for any write is the time for the initiator's ABC-send to be received by the initiator, which is at most h.

The remainder of the proof is devoted to showing sequential consistency. Fix admissible execution σ.

Define the sequence of operations τ as follows. Order the writes in σ in Abcast order. Now we explain where to insert the reads. We proceed in order from the beginning of σ. [$\text{Read}_p(X)$, $\text{Ret}_p(X, v)$] goes immediately after the latest of (1) the previous operation for p (either read or write, on any object), and (2) the write that spawned the latest update of p's local copy of X preceding the generation of the $\text{Ret}_p(X, v)$. (Break ties using process ids; e.g., if every process reads some object before any process writes any object, then τ begins with p_1's read, followed by p_2's read, etc.)

We must show $ops(\sigma)|p = \tau|p$ for all processes p. Fix some process p.

The relative ordering of two reads in $ops(\sigma)|p$ is the same in $\tau|p$ by definition of τ.

The relative ordering of two writes in $ops(\sigma)|p$ is the same in $\tau|p$ by Lemma 4.3.

Suppose in $ops(\sigma)|p$ that read R follows write W. By definition of τ, R comes after W in τ.

Suppose in $ops(\sigma)|p$ that read R precedes write W. Suppose in contradiction that R comes after W in τ. Then in σ there is some read $R' =$ [$\text{Read}_p(X)$, $\text{Ret}_p(X, v)$] and some write $W' =$ [$\text{Write}_q(X, v)$, $\text{Ack}_q(X)$] such that (1) R' equals R or occurs before R in σ, (2) W' equals W or follows W in the Abcast order, and (3) W' spawns the latest update to p's copy of X that precedes R''s read. But in σ, R' finishes before W starts. Since updates are performed in σ in Abcast order (Lemma 4.3), R' cannot see W''s update, a contradiction.

We must show τ is legal. Consider read $R =$ [$\text{Read}_p(X)$, $\text{Ret}_p(X, v)$] in τ. Let W be the write in σ that spawns the latest update to p's copy of X preceding R's read of p's copy of X. Clearly $W =$ [$\text{Write}_q(X, v)$, $\text{Ack}_q(X)$] for some q. (If there is no such W, then consider an imaginary write at the beginning of σ.) By the definition of τ, R follows W in τ. We must show that no other write to X falls in between W and R in τ. Suppose in contradiction that $W' =$ [$\text{Write}_r(X, w)$, $\text{Ack}_r(X)$] does. Then by Lemma 4.3, the update for W' follows the update for W at every process in σ.

Case 1: $r = p$. Since τ preserves the order of operations at p, W' precedes R in σ. Since the update for W' follows the update for W in σ, R sees W''s update, not W's, contradicting the choice of W.

Case 2: $r \neq p$. By definition of τ, there is some operation in $ops(\sigma)|p$ that, in τ, precedes R and follows W' (otherwise R would not follow W'). Let O be the first such operation.

Suppose O is a write to some object Y. By Lemma 4.4, O's update to p's copy of Y precedes R's read of p's copy of X. Since updates are done in Abcast order, the update for W' occurs at p before the update for O, and thus before R's read, contradicting the choice of W.

Suppose O is a read. By the definition of τ, O is a read of X, and W''s update to p's copy of X is the latest one preceding O's read (otherwise O would not follow W'). Since updates are done in Abcast order, the value from W' supersedes the value from W, contradicting the choice of W. ■

Theorem 4.1 implies that this algorithm does not guarantee linearizability. We can also explicitly construct an admissible execution that violates linearizability as follows. The initial value of X is 0. Process p writes 1 to X. The ABC-send for the write occurs at time t. It arrives at process r at time t and at process q at time $t + h$. Meanwhile, r performs a read at time t and gets the new value 1, while q performs a read at time $t + h/2$ and gets the old value 0. No permutation of these operations can both conform to the read/write specification and preserve the relative real-time orderings of all non-overlapping operations.

Fast Writes

We now discuss the algorithm that ensures sequential consistency with fast writes (time 0) and slow reads (time at most h). When a Read(X) comes in to p, if p has no pending updates (to any object, not just X) that it initiated, then it Returns the current value of its copy of X. Otherwise, it waits for all pending writes to complete and then returns. This is done by maintaining a count of the pending writes and waiting for it to be zero. When a Write(X) comes in to p, it is handled very similarly to the other algorithm; however, it is Acked immediately. Effectively, the algorithm pipelines write updates generated at the same process.

Specifically, the state of each process consists of the following variables:
num : integer, initially 0 (number of pending updates initiated by this process),
copy of every object, initially equal to its initial value.

The transition function of process p appears in Fig. 5.

Theorem 4.6 *There exists a sequentially consistent implementation of read/write objects with* $|Read| = h$ *and* $|Write| = 0$.

Proof: Consider the algorithm just presented. Clearly every write takes 0 time. The worst-case time for a read occurs if the return must wait for the initiator to receive its own ABC-send for a pending write. This takes at most h time.

```
Read_p(X):
  if num = 0 then
      generate Ret_p(X, v), where v is the value of p's copy of X
  endif

Write_p(X, v):
  num := num + 1
  ABC-send(X, v)
  generate Ack_p(X)

ABC-receive(X, v, i) from q:
  set local copy of X to v
  if p = q then
      num := num −1
      if num = 0 then
          generate Ret_p(X, v), where v is the value of p's copy of X
      endif
  endif
```

Figure 5: Sequentially consistent fast write algorithm.

The structure of the proof of sequential consistency is identical to that in the proof of Theorem 4.5. We just need a new proof for Lemma 4.4.

Lemma 4.4 is still true for this algorithm because when a Read occurs at p, if any update initiated by p is still waiting, then the Return is delayed until the latest such update is performed. ■

Theorem 4.2 implies that this algorithm does not guarantee linearizability. We can also construct an explicit scenario.

4.2 FIFO Queues

We show in Section 4.2.1 that in any linearizable implementation of a FIFO queue, the worst-case response time of an enqueue operation must depend on u. We then present in Section 4.2.2 a sequentially consistent implementation in which enqueue operations return instantaneously while the worst-case response time for a dequeue operation is h.

4.2.1 Lower Bound for Linearizability

We show that in any linearizable implementation of a FIFO queue the worst-case time for an enqueue is $u/2$ (assuming that at least two processes can enqueue to the same FIFO queue). The proof uses the technique of shifting, described in Section 2.4.

Theorem 4.7 *If Q is a FIFO queue with at least two enqueuers, then any linearizable implementation of Q must have $|Enq(Q)| \geq \frac{u}{2}$.*

Proof: Let p and q be two processes that can enqueue to Q and r be a process that dequeues from Q. Assume in contradiction that there is an implementation with $|Enq(Q)| < \frac{u}{2}$. Initially, Q is empty. By the specification of Q, there is an admissible execution α such that

- $ops(\alpha)$ is $\mathrm{Enq}_p(Q,1)\ \mathrm{Ack}_p(Q)\ \mathrm{Enq}_q(Q,2)\ \mathrm{Ack}_q(Q)\ \mathrm{Deq}_r(Q)\ \mathrm{Ret}_r(Q,1)$;
- $\mathrm{Enq}_p(Q,1)$ occurs at time 0, $\mathrm{Enq}_q(Q,2)$ occurs at time $\frac{u}{2}$, and $\mathrm{Deq}_r(Q)$ occurs at time u; and
- the message delays in α are d from p to q, $d-u$ from q to p, and $d-\frac{u}{2}$ for all other ordered pairs of processes.

Let $\beta = shift(shift(\alpha, p, -\frac{u}{2}), q, \frac{u}{2})$; i.e., we shift p later by $\frac{u}{2}$ and q earlier by $\frac{u}{2}$. The result is still an admissible execution, since by Lemma 2.3 the delay of a message from p or to q becomes $d-u$, the delay of a message from q or to p becomes d, and all other delays are unchanged. But $ops(\beta)$ is

$$\mathrm{Enq}_q(Q,2)\ \mathrm{Ack}_q(Q)\ \mathrm{Enq}_p(Q,1)\ \mathrm{Ack}_p(Q)\ \mathrm{Deq}_r(Q)\ \mathrm{Ret}_r(Q,1)\ .$$

which violates linearizability, because r's dequeue should return 2, not 1 (by the FIFO property). ■

The assumption about the number of enqueuers made in Theorem 4.7 is crucial to the results, since it can be shown that the algorithm of Theorem 3.6 is correct if there is only one enqueuer.

4.2.2 Upper Bound for Sequential Consistency

Informally, the algorithm works as follow. Each process keeps a local copy of every object, When a request to enqueue v to Q comes in to p, p broadcasts an update message with the object name, the operation name, and the value to be enqueued to all processes. The operation returns immediately. When a request to dequeue from Q comes in to p, p broadcasts an update message with the object name and the operation name. It does not generate a response.

```
$\text{Enq}_p(Q, v)$:
  ABC-send($Q, v$,"enq")
  generate $\text{Ack}_p(Q)$

$\text{Deq}_p(Q)$:
  ABC-send($Q$,"deq")

ABC-receive($Q, v$,"enq") from $q$:
  enqueue $v$ on local copy of $Q$

ABC-receive($Q$,"deq") from $q$:
  val := dequeue local copy of $Q$
  if $p = q$ then generate $\text{Ret}_p(Q,\text{val})$ endif
```

Figure 6: Sequentially consistent fast enqueue algorithm.

When an update message (either "dequeue" or "enqueue") is delivered to a process it handles it by performing the appropriate change (enqueue or dequeue) to the local copy of the object. If the update is a dequeue by the same process, the dequeue operation that is currently waiting returns the value that was dequeued from the local copy. (Note that by well-formedness, there is only one pending dequeue operation for a given process.)

In more detail, the state of each process consists of the following variables:
copy of every object, initially equal to its initial value
val : value (of a queue element)

The transition function of process p appears in Fig. 6.

To prove correctness of the algorithm we show:

Lemma 4.8 *In every admissible execution, all updates are done exactly once at each local copy, updates are done in the same order at each process, and this order preserves the per-process order.*

Theorem 4.9 *There exists a sequentially consistent implementation of FIFO queues with* $|Enq| = 0$ *and* $|Deq| = h$.

Proof: Consider the algorithm just presented. Clearly, the time for an enqueue is 0 and the time for a dequeue is at most h.

The remainder of the proof is devoted to showing sequential consistency. Fix some admissible execution σ.

Define the sequence of operations τ as follows: Order the operations in σ by Abcast order. From Lemma 4.8 it follows that operations by p are ordered in τ as they were ordered in σ, and thus $ops(\sigma)|p = \tau|p$, for all processes p.

It remains to show that τ is legal, i.e., that for every FIFO queue Q, $\tau|Q$ is in the serial specification of Q. Pick any Q and consider $\tau|Q = op_1 op_2 \ldots$. Suppose op_i is $[\mathrm{Deq}_p(Q), \mathrm{Ret}_p(Q, v)]$. Since the local updates at p occur in Abcast order (Lemma 4.8), updates at p to the local copy of Q occur in the same order as in τ, and the claim follows. ■

Theorem 4.7 implies that this algorithm does not guarantee linearizability. It is also possible to construct an explicit scenario which violates linearizability.

4.3 Stacks

These results are analogous to those for FIFO queues with Pop in place of Deq and Push in place of Enq.

Theorem 4.10 *If S is a stack with at least two pushers, then for any linearizable implementation of S, $|Push(S)| \geq \frac{u}{2}$.*

Theorem 4.11 *There exists a sequentially consistent implementation of stacks with $|Push| = 0$ and $|Pop| = h$.*

5 Conclusions and Further Research

The impact of the correctness guarantee on the efficiency of supporting it was studied under various timing assumptions. Although we still do not have a complete picture of this problem, our results indicate that supporting sequential consistency can be more cost-effective than supporting linearizability, for certain object types and under certain timing assumptions. Two other conclusions can be drawn from our results. First, perfect clocks admit more efficient implementations, and thus it may be worthwhile to provide such clocks. Second, knowing in advance the sharing patterns of the objects (i.e., how many processes access each one with a particular operation) results in faster implementations. Thus, the mcs can benefit from having the application program (the user) supply "hints" about the sharing patterns of the object.

Our work leaves open many interesting questions. Obviously, it is desirable to narrow the gaps between our upper and lower bounds. (Some recent results in this direction appear in [29].) It will be interesting to understand how practical issues such as local memory size and clock drift influence the bounds. It will be very interesting to obtain bounds on the response time of implementing other objects, e.g., Test&Set registers, under sequential consistency and linearizability. The cost measure we have chosen to analyze is response time, but there are other relevant measures, including throughput and network congestion.

The modular usage of atomic broadcast in our implementations of sequential consistency admits several extensions. For example, a bus provides an easy mechanism for atomic broadcast, and it might be possible to improve and simplify the correctness proof of the algorithm in [3] using this observation. Also, atomic broadcast algorithms can be made *fault-tolerant*. This can help in the design of memory consistency systems that can sustain failures of some of the processes. In general, the issue of fault-tolerance is rarely addressed in the current research on memory consistency. As multiprocessors scale up and the probability of failure increase, this will become an important concern.

The problem that we have studied is closely related to the problem of designing cache consistency schemes in which some sort of global ordering must be imposed on the operations ([11, 12, 14, 18, 24]). Our results show that making the definitions of these orderings more precise is important since seemingly minor differences in the definitions result in significant differences in the inherent efficiency of implementing them. Recently, several non-global conditions that are weaker than sequential consistency have been suggested, e.g., weak ordering ([17, 9, 1]), release consistency ([20]), pipelined memory ([27]), slow memory ([23]), causal memory ([4]), loosely coherent memory ([7]), and the definitions in [14] and [33]. It would be interesting to investigate the inherent efficiency of supporting these consistency guarantees. In order to do so, crisp and precise definitions of these conditions are needed.

It is clear that efficiency, in general, and response time, in particular, are not the only criteria for evaluating consistency guarantees. In particular, the ease of designing, verifying, programming, and debugging algorithms using such shared memories is very important.

As multiprocessor systems become larger, distributed implementations of shared virtual memory are becoming more common. (Truly shared memories, or even buses, cannot be used in systems with a large number of processors.) Such implementations and their evaluation relate issues concerning multiprocessor architecture, programming language design, software engineering, and the theory of concurrent systems. We hope our work contributes toward a more solid ground for this interaction.

Acknowledgements: The authors thank Sarita Adve, Roy Friedman, Mark Hill, and Rick Zucker for helpful comments on an earlier version of this paper. We especially thank Martha Kosa for a careful reading.

References

[1] S. Adve and M. Hill, "Weak Ordering—A New Definition," *Proc. 17th Int. Symp. on Computer Architecture,* 1990, pp. 2–14.

[2] S. Adve and M. Hill, "Implementing Sequential Consistency in Cache-Based Systems," *Proc. Int. Conf. on Parallel Processing,* 1990, pp. I-47–50.

[3] Y. Afek, G. Brown, and M. Merritt. "A Lazy Cache Algorithm," *Proc. 1st ACM Symp. on Parallel Algorithms and Architectures,* 1989, pp. 209–222.

[4] M. Ahamad, P. Hutto, and R. John, *Implementing and Programming Causal Distributed Shared Memory,* TR GIT-CC-90-49, Georgia Inst. of Tech., December 1990.

[5] H. Attiya, "Implementing FIFO Queues and Stacks," to appear in *Proc. Int. Workshop on Distributed Algorithms,* October 1991.

[6] H. Attiya and J. L. Welch, "Sequential Consistency versus Linearizability," *Proc. 3rd ACM Symp. on Parallel Algorithms and Architectures,* 1991, pp. 304–315.

[7] J. Bennett, J. Carter, and W. Zwaenepoel, "Munin: Distributed Shared Memory Based on Type-Specific Memory Coherence," *Proc. 2nd ACM Symp. on Principles and Practice of Parallel Processing,* 1990, pp. 168–176.

[8] P. Bernstein, V. Hadzilacos, and H. Goodman, *Concurrency Control and Recovery in Database Systems,* Addison-Wesley, Reading, MA, 1987.

[9] R. Bisiani, A. Nowatzyk, and M. Ravishankar, "Coherent Shared Memory on a Distributed Memory Machine," *Proc. Int. Conf. on Parallel Processing,* 1989, pp. I-133–141.

[10] K. Birman and T. Joseph, "Reliable Communication in the Presence of Failures," *ACM Trans. on Computer Systems,* vol. 5, no. 1, pp. 47–76.

[11] W. Brantley, K. McAuliffe, and J. Weiss, "RP3 Processor-Memory Element," *Proc. Int. Conf. on Parallel Processing,* 1985, pp. 782–789.

[12] L. M. Censier and P. Feautrier, "A New Solution to Coherence Problems in Multicache Systems," *IEEE Trans. on Computers,* vol. C-27, no. 12, pp. 1112–1118.

[13] J. Chang and N. F. Maxemchuk, "Reliable Broadcast Protocols," *ACM Trans. on Computer Systems,* vol. 2, no. 3, pp. 251-273.

[14] W. W. Collier, "Architectures for Systems of Parallel Processes," IBM TR 00.3253, Poughkeepsie, NY, January 1984.

[15] E. W. Dijkstra, "Hierarchical Ordering Of Sequential Processes," *Acta Informatica,* 1971, pp. 115–138.

[16] M. Dubois and C. Scheurich, "Memory Access Dependencies in Shared-Memory Multiprocessors", *IEEE Trans. on Software Engineering,* vol. 16, no. 6 (June 1990), pp. 660–673.

[17] M. Dubois, C. Scheurich, and F. A. Briggs, "Memory Access Buffering in Multiprocessors," *Proc. 13th Int. Symp. on Computer Architecture,* June 1986, pp. 434–442.

[18] M. Dubois, C. Scheurich, and F. A. Briggs, "Synchronization, Coherence and Event Ordering in Multiprocessors," *IEEE Computer,* vol. 21, no. 2, pp. 9–21.

[19] H. Garcia-Molina and A. Spauster, "Message Ordering in a Multicast Environment," *Proc. Int. Conf. on Distributed Computing Systems,* 1989, pp. 354-361.

[20] K. Gharachorloo, D. Lenoski, J. Laudon, P. Gibbons, A. Gupta and J. Hennessey, "Memory Consistency and Event Ordering in Scalable Shared-Memory Multiprocessors," *Proc. 17th Int. Symp. on Computer Architecture,* 1990, pp. 15-26.

[21] M. Herlihy, "Wait-Free Implementations of Concurrent Objects," *Proc. ACM Symp. on Principles of Distributed Computing,* 1988, pp. 276–290.

[22] M. Herlihy and J. Wing, "Linearizability: A Correctness Condition for Concurrent Objects," *ACM Trans. on Programming Languages and Systems,* vol. 12, no. 3, pp. 463–492.

[23] P. Hutto and M. Ahamad, *Slow Memory: Weakening Consistency to Enhance Concurrency in Distributed Shared Memories,* TR GIT-ICS-89/39, Georgia Inst. of Tech., October 1989.

[24] L. Lamport, "How to Make a Multiprocessor Computer that Correctly Executes Multiprocess Programs," *IEEE Trans. on Computers,* vol. C-28, no. 9, pp. 690–691.

[25] L. Lamport, "On Interprocess Communication. Parts I and II," *Distributed Computing,* vol. 1, no. 2 (1986), pp. 77-101.

[26] K. Li and P. Hudak, "Memory Coherence in Shared Virtual Memory Systems," *ACM Trans. on Computer Systems,* vol. 7, no. 4, pp. 321–359.

[27] R. Lipton and J. Sandberg, *PRAM: A Scalable Shared Memory,* TR CS-TR-180-88, Princeton University, September 1988.

[28] J. Lundelius and N. Lynch, "An Upper and Lower Bound for Clock Synchronization," *Information and Control,* vol. 62, nos. 2/3, pp. 190–204.

[29] M. Mavronicolas and D. Roth, "On the Costs of Sequential Consistency and Linearizability," in preparation.

[30] S. Min and J. Baer, "A Timestamp-Based Cache Coherence Scheme," *Proc. Int. Conf. on Parallel Processing,* 1989, pp. I-23–32.

[31] J. Misra, "Axioms for Memory Access in Asynchronous Hardware Systems," *ACM Trans. on Programming Languages and Systems,* vol. 8, no. 1, pp. 142–153.

[32] C. Papadimitriou, *The Theory of Concurrency Control,* Computer Science Press, Rockville, MD, 1986.

[33] U. Ramachandran, M. Ahamad, and M. Y. Khalidi, "Coherence of Distributed Shared Memory: Unifying Synchronization and Data Transfer," *Proc. Int. Conf. on Parallel Processing,* 1989, pp. II-160–169.

[34] C. Scheurich and M. Dubois, "Correct Memory Operation of Cache-Based Multiprocessors," *Proc. 14th Int. Symp. on Computer Architecture,* 1987, pp. 234–243.

A Atomic Broadcast

The atomic broadcast algorithm employed by our algorithms is based on assigning timestamps to messages. Each process maintains a local timestamp (counter) and a vector with (conservative) estimates of the timestamps of all other processes. A process keeps a timestamp bigger than or equal to the timestamps of all the other processes (according to its estimates). Upon a request to broadcast a message, the message is tagged with the requester's current timestamp. Each process maintains a set of messages that are waiting to be delivered. A message with timestamp x is delivered only when the process is certain that all other messages with timestamp $\leq x$ have arrived at it. This is done by waiting to learn that all processes have increased their timestamp to be at least $x + 1$.[9] Once it learns that all processes have increased their timestamps beyond x, the process handles all pending messages with timestamps less than or equal to x, in order, breaking ties using process ids.

More precisely, to broadcast a message m, p sends a message (t_p,m) to all processes (including itself), where t_p is p's current timestamp. It then increases its own timestamp by one, and returns. When a process q receives a message with timestamp t_p from p, it saves it in a list of pending messages, sorted by timestamp and process id. It then increases its timestamp to be at least as large as $t_p + 1$ and sends a timestamp increase message "timestamp(t_q, q)".

When a process receives a timestamp increase message, it updates the timestamp entry for the sender, and checks to see if there are any pending messages whose timestamp is strictly less than all processes' timestamps (saved in its local vector). These messages are delivered in increasing timestamp order, breaking ties using process ids.

The algorithm uses the following data types:

```
        timestamp = integer
        message = record with fields
                mess : string (message to be delivered)
                ts : timestamp (assigned by initiator)
                id : process id (id of initiator)
```

Each process knows n, the total number of processes.
The state of each process consists of the following components:

```
        ts : array[1..n] of integer, all initially 0
           (estimate (from below) timestamps of all processes)
        pending : set of message, initially empty
           (set of message waiting to be delivered)
```

The transition function of process p appears in Fig. 7.

To show that this algorithm implements atomic broadcast, we must show, for any admissible execution, that messages are delivered at the same order to all processes. The ordering of

[9] For simplicity, the algorithm presented here assumes FIFO channels. This assumption can be removed if sequence numbers are employed.

```
ABC-send$_p(m)$:
  send (ts[$p$],$m$) to all processes
  ts[$p$] := ts[$p$] + 1

receive ($t$,$m$) from $q$:
  add ($m$,$t$,$q$) to pending
  if $t + 1 >$ ts[$p$] then
        ts[$p$] := $t + 1$
        send timestamp(ts[$p$]) to all processes
  endif

receive timestamp($t$) from $q$:
  ts[$q$] := $t$
  repeat
        let $E$ be element with smallest (ts,id) pair in pending
        if for some $q$, ts[$q$] $\leq$ $E$.ts then exit
        deliver $E.m$ { this is the ABC-receive }
        remove $E$ from pending
  endrepeat
```

Figure 7: Atomic broadcast algorithm.

messages is done by timestamps (breaking ties with process ids). The resulting sequence respects the order at each process by construction and because of the way timestamps are assigned.

More formally, fix some admissible execution σ of the algorithm. The next lemma follows immediately from the code.

Lemma A.1 *Let p be any process. Then every message broadcast by p in σ is given a unique timestamp in increasing order.*

This immediately implies:

Lemma A.2 *The timestamps assigned to messages in σ, together with process ids, form a total order.*

This total order is called *timestamp order.*

Lemma A.3 *Let p be any process. Then all messages are delivered to p in σ in timestamp order.*

Proof: Let (t_1, q_1) be the timestamp of the message m_1, and let (t_2, q_2) be the timestamp of the message m_2. Suppose, by way of contradiction, that $(t_1, q_1) < (t_2, q_2)$ but m_2 was delivered to p before m_1.

When m_2 is delivered to p, it cannot yet have the message m_1 in pending, because otherwise it would deliver it before m_2. By the code, in order to deliver m_2, it must be that $ts_p[q_1] > t_2$. But then p must have received a timestamp message from q_1 with a timestamp $t \geq t_2 + 1$. Since $(t_1, q_1) < (t_2, q_2)$ it must be that $t_1 \leq t_2$, and hence $t > t_1$. By the code, the message m_1 was sent before the timestamp message. But then the FIFO property of the communication system implies that p has already received m_1. A contradiction. ■

The next lemma guarantees that each message is delivered within time $2d$ from the initiation of the operation.

Lemma A.4 *If process p broadcasts a message m, then m is delivered at each process within time at most $2d$ in σ.*

Proof: Assume p broadcasts m at time T, with timestamp x. By time $T+d$ all processes will get the message (x, m), and will set their timestamps to be at least $x+1$, sending a timestamp increase message to all other processes, if necessary. Thus, by time $T + 2d$, all processes will have in their timestamp vectors values that are strictly larger than x, and will deliver m. ■

Lemmas A.3 and A.4 prove the following theorem.

Theorem A.5 *The algorithm in Fig. 7 is an atomic broadcast algorithm with $h = 2d$.*