Lightweight Causal and Atomic Group Multicast*

Kenneth Birman
Andre Schiper
Pat Stephenson

TR 91-1192
February 1991

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This work was supported by the Defense Advanced Research Projects Agency (DoD) under DARPA/NASA subcontract NAG2-593 administered by the NASA Ames Research Center, and by grants from GTE, IBM, and Siemens, Inc. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.
Lightweight Causal and Atomic Group Multicast

Kenneth Birman  
Department of Computer Science, Cornell University

Andre Schiper  
Ecole Polytechnique Federale de Lausanne, Switzerland

Pat Stephenson  
Department of Computer Science, Cornell University

February 21, 1991

Abstract

The Isis toolkit is a distributed programming environment based on support for virtually synchronous process groups and group communication. We present a new suite of protocols in support of this model. Our approach revolves around a multicast primitive, called CBCAST, which implements a fault-tolerant, causally ordered message delivery. This primitive can be used directly, or extended into a totally ordered multicast primitive, called ABCAST. It normally delivers messages immediately upon reception, and imposes a space overhead proportional to the size of the groups to which the sender belongs, usually a small number. We conclude that process groups and group communication can achieve performance and scaling comparable to that of a raw message transport layer - a finding contradicting the widespread concern that this style of distributed computing may be unacceptably costly.

Keywords and phrases: Distributed computing, process groups, fault-tolerance, scalability, reliable multicast, atomic broadcast, virtual synchrony, ABCAST, CBCAST, Isis.

*This work was supported by the Defense Advanced Research Projects Agency (DoD) under DARPA/NASA subcontract NAG2-593 administered by the NASA Ames Research Center, and by grants from GTE, IBM, and Siemens, Inc. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.
1 Introduction

1.1 The ISIS Toolkit

The Isis Toolkit [BJKS88] provides a variety of tools for building software in loosely coupled distributed environments. The system has been successful in addressing problems of distributed consistency, cooperative distributed algorithms, and fault-tolerance. At the time of this writing, Version 2.1 of the Toolkit was in use at several hundred locations worldwide.

Two aspects of Isis are key to its overall approach:

- An implementation of virtually synchronous process groups. Such a group consists of a set of processes cooperating to execute a distributed algorithm, replicate data, provide a service fault-tolerantly, or otherwise exploit distribution.

- A collection of reliable multicast protocols with which processes and group members interact with groups. Reliability in Isis encompasses failure atomicity, delivery ordering guarantees, and a form of group addressing atomicity, under which membership changes are synchronized with group communication.

Although Isis supports a wide range of multicast protocols, a protocol called CBCAST accounts for the majority of communication in the system. In fact, many of the Isis tools are little more than invocations of this communication primitive. For example, the Isis replicated data tool uses a single (asynchronous) CBCAST to perform each update and locking operation; reads require no communication at all. A consequence is that the cost of CBCAST represents the dominant performance bottleneck in the Isis system.

The original Isis CBCAST protocol was costly in part for structural reasons and in part because of the protocol used [BJ87b]. The implementation was within a protocol server, hence all CBCAST communication was via an indirect path. Independent of the cost of the protocol itself, this indirection was expensive. Furthermore, the protocol server proved difficult to scale, limiting the initial versions of Isis to networks of a few hundred nodes. With respect to the protocol used, our initial implementation favored generality over specialization, permitting extremely flexible destination addressing. It used a piggybacking algorithm that achieved the CBCAST ordering property but required periodic garbage collection.

The case for flexibility in addressing seems weaker today. Experience with Isis has left us with substantial insight into how the system is used, permitting us to focus on core functionality.
protocols described in this paper support highly concurrent applications, scale to systems with large numbers of potentially overlapping process groups, and bound the overhead associated with piggybacked information in proportion to the size of the process groups to which the sender of a message belongs. Although slightly less general than the earlier solution, the new protocols are able to support the Isis toolkit and all Isis applications with which we are familiar. The benefit of this reduction in generality has been a substantial increase in the performance and scalability of our system. In fact, the new protocol suite has no evident limits to the scale of system it could support. In the common case of an application with localized, bursty communication, most multicasts will carry only a small overhead regardless of the size or number of groups used, and a message will be delayed only if it actually arrives out of order.

The paper is structured as follows. Section 2 discusses the types of process groups supported by Isis and the patterns of group usage and communication that have been observed among current Isis applications. Section 3 surveys prior work on multicast. Section 4 formalizes the virtually synchronous multicasting problem and the properties that a CBCAST or ABCAST protocol must satisfy. Sections 5 introduces our new technique in a single process group; multiple groups are considered in Section 6. Section 7 considers a number of Isis-specific implementation issues. The paper concludes with a discussion of the performance of our initial implementation, in Section 8.

2 Experience with Isis users

We begin by reviewing the types of groups and patterns of group usage seen in existing Isis applications. This material is discussed in more detail in [BC90].

Isis supports four types of groups, illustrated in Figure 1. The simplest of these is denoted the peer group. In a peer group, processes cooperate as equals in order to get a task done. They may manage replicated data, subdivide tasks, monitor one another's status, or otherwise engage in a closely coordinated distributed action. Another common structure is the client/server group. Here, a peer group of processes act as servers on behalf of a potentially large set of clients. Clients interact with the servers in a request/reply style, either by picking a favorite server and issuing RPC calls to it, or by multicasting to the whole server group. In the later case, servers will often multicast their replies both to the appropriate client and to one another. A diffusion group is a type of client-server group in which the servers broadcast messages to the full set of servers and clients. Clients are passive and simply receive messages. Diffusion groups arise in any application that broadcasts information to large number of sites, for example on a brokerage trading floor. Finally, hierarchical group structures arise when large server groups are needed in a distributed system [CB89,GMS89]. Hierarchical groups are tree-structured sets of groups. A root group maps the initial connection
request to an appropriate subgroup, and the application subsequently interacts only with this subgroup. Data is partitioned among the subgroups, and although a large-group communication mechanism is available, it is rarely needed.

Many Isis applications use more than one of these structures, employing overlapping groups when mixed functionality is desired. For example, a diffusion group used to disseminate stock quotes would almost always be overlaid by a client-server group through which brokerage programs register their interest in specific stocks. Nonetheless, existing Isis applications rarely use large numbers of groups. Groups change membership infrequently, and generally contain just enough members for fault-tolerance or load-sharing (e.g., 3-5 processes). On the other hand, the number of clients of a client/server or diffusion group may be large (hundreds).

Through studies of Isis users [BC90,BCG91] we have concluded that these patterns are in part artifacts of the way Isis evolved. In versions of Isis prior to the one discussed here, groups were fairly heavy-weight entities. Applications obtained acceptable performance only by ensuring that communication to a group was much more frequent than membership changes. Looking to the future, we expect our system to continue supporting these four types of groups. We also expect that groups will remain small, (except for the client set of a client-server or diffusion group). However,
as we rebuild Isis around the protocols described here and move the key modules into lower layers of the operating system, groups and group communication can be expected to get much cheaper. These costs seem to be a dominant factor preventing Isis users from employing very large numbers of groups, especially in cases where process groups naturally model some sort of application-level data type or object. As a result, we expect that for some applications, groups will substantially outnumber processes. Furthermore, groups may become much more dynamic, because the cost of joining or leaving a group can be substantially reduced using the protocols developed in this paper.

To illustrate these points, we consider some applications that would have these characteristics. A scientific simulation employing an n-dimensional grid might use a process group to represent the neighbors of each grid element. A network information service running on hundreds of sites might replicate individual data items using small process groups; the result would be a large group containing many smaller data replication domains, perhaps moving data in response to access patterns. Similarly, a process group could be used to implement replicated objects in a modular application that imports many such objects. In each case, the number of process groups would be huge and the overlap between groups extensive.

The desire to support applications like these represents a primary motivation for the research reported here. The earlier Isis protocols have proven inflexible and difficult to scale, and it seems unlikely that they could be used to support the highly-dynamic, large-scale applications that now interest us. The protocols reported here respond to these new needs, enabling the exploration of such issues as support for parallel processing, the use of multicast communication hardware, mechanisms to enforce realtime deadlines and message priorities, etc.

3 Prior work on group communication protocols

Our communication protocols evolved from a causal message delivery protocol developed by Schiper [SES...]. and are based on work by Fidge and Mattern [Fid88,Mat89]. In the case of a single process group, the algorithm resembles protocols developed by Ladin [LLS90] and Peterson [PBS89]. However, our work generalizes these protocols in the following respects:

- Both of the other multicast protocols address causality only in the context of a single process group. Our solution transparently addresses the case of multiple, overlapping groups. In [BCG91] we argue that a multicast protocol must respect causality to be used asynchronously (without blocking the sender until remote delivery occurs). Asynchronous communication is the key to high performance in group-structured distributed applications, and is a central feature of Isis.
The Isis architecture treats client/server groups and diffusion groups as sets of overlaid groups, and optimizes the management of causality information for this case. Both the clients and servers can multicast directly and fault-tolerantly within these subgroups of a client/server group. Peterson's protocols do not support these styles of group use and communication. Ladin's protocol supports client/server interactions, but not diffusion groups, and does not permit clients to multicast directly to server groups.

- Ladin's protocol uses stable storage as part of the fault-tolerance method. Our protocol uses a notion of message *stability* that requires no external storage.

Our CBCAST protocol can be extended to provide a total message delivery ordering, inviting comparison with atomic broadcast (ABCAST) protocols [CM84,BJ87b,GMS89,PGM85,VRB89]. Again, the extensions supporting multiple groups represent our primary contribution. However, our ABCAST protocol also uses a delivery order consistent with causally, permitting it to be used asynchronously. A delivery ordering might be total without being causal, and indeed, several of the protocols cited would not provide this guarantee.

## 4 Execution model

We now formalize the model and the problem to be solved.

### 4.1 Basic system model

The system is composed of processes $P = \{p_1, p_2, ..., p_n\}$ with disjoint memory spaces. Initially, we assume that this set is static and known in advance; later we relax this assumption. Processes fail by crashing detectably (a *fail-stop* assumption); notification is provided by a failure detection mechanism, described below. When multiple processes need to cooperate, e.g. to manage replicated data, subdivide a computation, monitor one another's state, and so forth, they can be structured into *process groups*. The set of such groups is denoted by $G = \{g_1, g_2, ...\}$.

Each process group has a name and a set of member processes. Members join and leave dynamically; a failure causes a departure from all groups to which a process belongs. The members of a process group need not be identical, nor is there any limit on the number of groups to which a process may belong. The protocols presented below all assume that processes only multicast to groups that they are members of, and that all multicasts are directed to the full membership of a single group. (We discuss client/server groups in Section 7.)
Our system model is unusual in assuming an external service that implements the process group abstraction. This accurately reflects our current implementation, which obtains group membership management from a pre-existing Isis process-group server. In fact, however, this requirement can be eliminated, as discussed in Section 7.4.

The interface by which a process joins and leaves a process group will not concern us here, but the manner in which the group service communicates membership information to a process is relevant. A view of a process group is a list of its members. A view sequence for \( g \) is a list \( \text{view}_0(g), \text{view}_1(g), \ldots, \text{view}_n(g) \), where

1. \( \text{view}_0(g) = \emptyset \).
2. \( \forall i : \text{view}_i(g) \subseteq P \), where \( P \) is the set of all processes in the system.
3. \( \text{view}_i(g) \) and \( \text{view}_{i+1}(g) \) differ by the addition or subtraction of exactly one process.

Processes learn of the failure of other group members only through this view mechanism, never through any sort of direct observation.

We assume that direct communication between processes is always possible; the software implementing this is called the message transport layer. Within our protocols, processes always communicate using point-to-point and multicast messages; the latter may be transmitted using multiple point-to-point messages if no more efficient alternative is available. The transport communication primitives must provide lossless, uncorrupted, sequenced message delivery. The message transport layer is also assumed to intercept and discard messages from a failed process once the failure detection has been made. This guards against the possibility that a process might hang for an extended period (e.g. waiting for a paging store to respond), but then attempt to resume communication with the system. Obviously, transient problems of this sort cannot be distinguished from permanent failures, hence there is little choice but to treat both the same way by forcing the faulty process to run a recovery protocol.

Our protocol architecture permits application builders to define new transport protocols, perhaps to take advantage of special hardware. The implementation described in this paper uses a transport that we built over an unreliable datagram layer, but also has an experimental protocol that runs over ethernet hardware multicast.

The execution of a process is a partially-ordered sequence of events, each corresponding to the execution of an indivisible action. An acyclic event order, denoted \( P \), reflects the dependence of events occurring at process \( p \) upon one another. The event \( \text{send}_p(m) \) denotes the transmission of \( m \) by process \( p \) to a set of one or more destinations, \( \text{dests}(m) \); the reception of message \( m \) by
process \( p \) is denoted \( \text{rcv}_p(m) \). We omit the subscript when the process is clear from the context. If \( |\text{dests}(m)| > 1 \) we will assume that \textit{send} puts messages into all communication channels in a single action that might be interrupted by failure, but not by other \textit{send} or \textit{rcv} actions.

We denote by \( \text{rcv}_p(\text{view}_i(g)) \) the event by which a process \( p \) belonging to \( g \) "learns" of \( \text{view}_i(g) \).

We distinguish the event of receiving a message from the event of delivery, since this allows us to model protocols that delay message delivery until some condition is satisfied. The delivery event is denoted \( \text{deliver}_p(m) \) where \( \text{rcv}_p(m) \prec \text{deliver}_p(m) \).

When a process belongs to multiple groups, we may need to indicate the group in which a message was sent, received, or delivered. We will do this by extending our notation with a second argument; for example, \( \text{deliver}_p(m, g) \), will indicate that message \( m \) was delivered at process \( p \), and was sent by some other process in group \( g \).

As in [Lam78], we define the potential causality relation for the system, \( \rightarrow \), as the transitive closure of the relation defined as follows:

1. If \( \exists p : e \rightarrow e' \), then \( e \rightarrow e' \)
2. \( \forall m : \text{send}(m) \rightarrow \text{rcv}(m) \)

For messages \( m \) and \( m' \), the notation \( m \rightarrow m' \) will be used as a shorthand for \( \text{send}(m) \rightarrow \text{send}(m') \).

Finally, for demonstrating liveness, we assume that any message sent by a process is eventually received unless the sender or destination fails, and that failures are detected and eventually reflected in new group views omitting the failed process.

### 4.2 Virtual synchrony properties required of multicast protocols

Earlier, we stated that Isis is a \textit{virtually synchronous} programming environment. Intuitively, this means that users can program as if the system scheduled one distributed event at a time (i.e. group membership changes, multicast, and failures). Were a system to actually behave this way, we would call it \textit{synchronous}; such an environment would greatly simplify the development of distributed algorithms but offers little opportunity to exploit concurrency. The "schedule" used by Isis is, however, synchronous in appearance only. The ordering requirements of the tools in the Isis toolkit have been analyzed, and the system actually enforces only the degree of synchronization needed in each case [BJ87a]. This results in what we call a \textit{virtually} synchronous execution, in which operations are often performed concurrently and multicasts are often issued asynchronously.
(without blocking), but algorithms can still be developed and reasoned about using a simple, synchronous model.

Virtual synchrony has two major aspects.

1. **Address expansion.** It should be possible to use group identifiers as the destination of a multicast. The protocol must expand a group address into a destination list and deliver the message such that:

   a. All the recipients are in identical group views when the message arrives.
   b. The destination list consists of precisely the members of that view.

   The effect of these rules is that the expansion of the destination list and message delivery appear as a single, instantaneous event.

2. **Delivery atomicity and order.** This involves delivery of messages fault-tolerantly (either all operational destinations eventually receive a message, or, and only if the sender fails, none do). Furthermore, when multiple destinations receive the same message, they observe consistent delivery orders, in one of the two senses detailed below.

Two types of delivery ordering will be of interest here. We define the causal delivery ordering for multicast messages \( m \) and \( m' \) as follows:

\[
m \rightarrow m' \implies \forall p \in \text{dests}(m) \cap \text{dests}(m') : \text{deliver}(m)^p \text{deliver}(m').
\]

CBCAST provides only the causal delivery ordering. If two CBCAST's are concurrent, the protocol places no constraints on their relative delivery ordering at overlapping destinations. ABCAST extends the causal ordering into a total one, by ordering concurrent messages \( m \) and \( m' \) such that:

\[
\exists m, m', p \in g : \text{deliver}_p(m, g)^p \text{deliver}_p(m', g) \implies \forall q \in g : \text{deliver}_q(m, g)^q \text{deliver}_q(m', g).
\]

Note that this definition of ABCAST only orders messages sent to the same group; other definitions are possible. We discuss this further in Section 6.2. Because the ABCAST protocol orders concurrent events, it is more costly than CBCAST, requiring synchronous solutions where the CBCAST protocol admits efficient, asynchronous solutions.

Although one can define other sorts of delivery orderings, our work on ISIS suggests that this is not necessary. The higher levels of the ISIS toolkit are themselves implemented almost entirely using
asynchronous CBCAST [BJ89,Sch88]. In fact, Schmuck shows in [Sch88] that many algorithms
specified in terms of ABCAST can be modified to use CBCAST without compromising correctness.
Further, he demonstrates that both protocols are complete for a class of delivery orderings. For
example, CBCAST can emulate any ordering property that permits message delivery on the first
round of communication.

Fault tolerance and message delivery ordering are not independent in our model. A process will not
receive further multicasts from a faulty sender after observing it to fail; this requires that multicasts
in progress at the time of the failure be flushed from the system before the view corresponding to the
failure can be delivered to group members. Furthermore, failures will not leave gaps in a causally-
related sequence of multicasts. That is, if $m \rightarrow m'$ and a process $p_i$ has received $m'$, it need not be
cconcerned that a failure could somehow prevent $m$ from being delivered to any of its destinations
(even if the destinations of $m$ and $m'$ don't overlap). Failure atomicity alone would not yield either
guarantee.

4.3 Vector time

Our delivery protocol is based on a type of logical clock called a vector clock. The vector time
protocol maintains sufficient information to represent $\rightarrow$ precisely.

A vector time for a process $p_i$, denoted $VT(p_i)$, is a vector of length $n$ (where $n = |P|$), indexed
by process-id.

1. When $p_i$ starts execution, $VT(p_i)$ is initialized to zeros.
2. For each event $send(m)$ at $p_i$, $VT(p_i)[i]$ is incremented by 1.
3. Each message multicast by process $p_i$ is timestamped with the incremented value of $VT(p_i)$.
4. When process $p_j$ delivers a message $m$ from $p_i$ containing $VT(m)$, $p_j$ modifies its vector clock
   in the following manner:

   $\forall k \in 1..n : VT(p_j)[k] = max(VT(p_j)[k], VT(m)[k])$

That is, the vector timestamp assigned to a message $m$ counts the number of messages, on a
per-sender basis, that causally precede $m$.

Rules for comparing vector timestamps are:

1. $VT_1 \leq VT_2$ iff $\forall i : VT_1[i] \leq VT_2[i]$
It can be shown that given messages \( m \) and \( m' \), \( m\rightarrow m' \) iff \( VT(m) < VT(m') \): vector timestamps represent causality precisely.

Vector times were proposed in this form by Fidge and Mattern [Fid88, Mat89]; the latter includes a good survey. Other researchers have also used vector times or similar mechanisms [WPE+83, LLS90, Sch88, Mar84]. As noted earlier, our work is an outgrowth of the protocol presented in [SES89], which uses vector times as the basis for a protocol that delivers point-to-point messages in an order consistent with causality.

5 The CBCAST and ABCAST protocol

This section presents our new CBCAST and ABCAST protocols. We initially consider the case of a single process group with fixed membership; multiple group issues are addressed in the next section. This section first introduces the causal delivery protocol, then extends it to a totally ordered ABCAST protocol, and finally considers view changes.

5.1 CBCAST protocol

Suppose that a set of processes \( P \) communicate using only broadcasts to the full set of processes in the system; that is, \( \forall m : \text{dests}(m) = P \). We now develop a delivery protocol by which each process \( p \) receives messages sent to it, delays them if necessary, and then delivers them in an order consistent with causality:

\[ m \rightarrow m' \implies \forall p : \text{deliver}_p(m) \rightarrow \text{deliver}_p(m') . \]

Our solution is derived using vector timestamps. The basic idea is to label each message with a timestamp, \( VT(m)[k] \), indicating precisely how many multicasts by process \( p_k \) precede \( m \). A recipient of \( m \) will delay \( m \) until \( VT(m)[k] \) messages have been delivered from \( p_k \). Since \( \rightarrow \) is an acyclic order accurately represented by the vector time, the resulting delivery order is causal and deadlock free.

The protocol is as follows:

1. Before sending \( m \), process \( p_i \) increments \( VT(p_i)[i] \) and timestamps \( m \).
2. On reception of message $m$ sent by $p_i$ and timestamped with $VT(m)$, process $p_j \neq p_i$ delays delivery of $m$ until:

$$\forall k : 1 \ldots n \left\{ \begin{array}{ll}
VT(m)[k] = VT(p_j)[k] + 1 & \text{if } k = i \\
VT(m)[k] \leq VT(p_j)[k] & \text{otherwise}
\end{array} \right.$$ 

Process $p_j$ need not delay messages received from itself. Delayed messages are maintained on a queue, the CBCAST delay queue. This queue is sorted by vector time, with concurrent messages ordered by time of receipt (however, the queue order will not be used until later in the paper).

3. When a message $m$ is delivered, $VT(p_j)$ is updated in accordance with the vector time protocol from Section 4.3.

Step 2 is the key to the protocol. This guarantees that any message $m'$ transmitted causally before $m$ (and hence with $VT(m') < VT(m)$) will be delivered at $p_j$ before $m$ is delivered. An example in which this rule is used to delay delivery of a message appears in Figure 2.

Figure 2: Using the VT rule to delay message delivery

We prove the correctness of the protocol in two stages. We first show that causality is never violated (safety) and then we demonstrate that the protocol never delays a message indefinitely (liveness).

**Safety.** Consider the actions of a process $p_j$ that receives two messages $m_1$ and $m_2$ such that $m_1 \rightarrow m_2$.

**Case 1.** $m_1$ and $m_2$ are both transmitted by the same process $p_i$. Recall that we assumed a lossless, live communication system, hence $p_j$ eventually receives both $m_1$ and $m_2$. By construction, $VT(m_1) < VT(m_2)$, hence under step 2, $m_2$ can only be delivered after $m_1$ has been delivered.
Case 2. \( m_1 \) and \( m_2 \) are transmitted by two distinct processes \( p_i \) and \( p'_i \). We will show by induction on the messages received by process \( p_j \) that \( m_2 \) cannot be delivered before \( m_1 \). Assume that \( m_1 \) has not been delivered and that \( p_j \) has received \( k \) messages.

Observe first that \( m_1 \rightarrow m_2 \), hence \( VT(m_1) < VT(m_2) \) (basic property of vector times). In particular, if we consider the field corresponding to process \( p_i \), the sender of \( m_1 \), we have

\[
VT(m_1)[i] \leq VT(m_2)[i]
\]

**Base case.** The first message delivered by \( p_j \) cannot be \( m_2 \). Recall that if no messages have been delivered to \( p_j \), then \( VT(p_j)[i] = 0 \). However, \( VT(m_1)[i] > 0 \) (because \( m_1 \) is sent by \( p_i \)), hence \( VT(m_2)[i] > 0 \). By application of step 2 of the protocol, \( m_2 \) cannot be delivered by \( p_j \).

**Inductive step.** Suppose \( p_j \) has received \( k \) messages, none of which is a message \( m \) such that \( m_1 \rightarrow m \). If \( m_1 \) has not yet been delivered, then

\[
VT(p_j)[i] < VT(m_1)[i]
\]

This follows because the only way to assign a value to \( VT(p_j)[i] \) greater than \( VT(m_1)[i] \) is to deliver a message from \( p_i \) that was sent subsequent to \( m_1 \), and such a message would be causally dependent on \( m_1 \). From relations 1 and 2 it follows that

\[
VT(p_j)[i] < VT(m_2)[i]
\]

By application of step 2 of the protocol, the \( k + 1 \)'st message delivered by \( p_j \) cannot be \( m_2 \). \( \square \)

**Liveness.** Suppose there exists a broadcast message \( m \) sent by process \( p_i \) that can never be delivered to process \( p_j \). Step 2 implies that either:

\[
\exists k : 1...n \left\{ \begin{array}{l} VT(m)[k] \neq VT(p_j)[k] + 1 \quad \text{for } k = i, \text{ or} \\ VT(m)[k] > VT(p_j)[k] \quad , k \neq i \end{array} \right. 
\]

and that \( m \) was not transmitted by process \( p_j \). We consider these cases in turn.

- \( VT(m)[i] \neq VT(p_j)[i] + 1 \); that is, \( m \) is not the next message to be delivered from \( p_i \) to \( p_j \). Notice that only a finite number of messages can precede \( m \). Since all messages are multicast to all processes and channels are lossless and sequenced, it follows that there must be some message \( m' \) sent by \( p_i \) that \( p_j \) received previously, has not yet delivered, and that is the next message from \( p_i \), i.e. \( VT(m')[i] = VT(p_j)[i] + 1 \). If \( m' \) is also delayed, it must be under the other case.
\( \exists k \neq i: VT(m)[k] > VT(p_j)[k] \). Let \( n = VT(m)[k] \). The \( n \)'th transmission of process \( p_k \)

must be some message \( m' \rightarrow m \) that has either not been received at \( p_j \), or was received and

is delayed. Under the hypothesis that all messages are sent to all processes, \( m' \) was already

multicast to \( p_j \). Since the communication system eventually delivers all messages, we may

assume that \( m' \) has been received by \( p_j \). The same reasoning that was applied to \( m \) can now

be applied to \( m' \). The number of messages that must be delivered before \( m \) is finite and \( > \) is

acyclic, hence this leads to a contradiction. \( \Box \)

5.2 Causal ABCAST protocol

The CBCAST protocol is readily extended into a causal, totally ordered, ABCAST protocol. We

should note that it is unusual for an ABCAST protocol to guarantee that the total order used

conforms with causality. For example, say that a process \( p \) asynchronously transmits message

\( m \) using ABCAST, then sends message \( m' \) using CBCAST, and that some recipient of \( m' \) now

sends \( m'' \) using ABCAST. Here we have \( m \rightarrow m' \rightarrow m'' \), but \( m \) and \( m'' \) are transmitted by different

processes. Many ABCAST protocols would use an arbitrary ordering in this case; our solution will

always deliver \( m \) before \( m'' \). This property is actually quite important: without it, few algorithms

could safely use ABCAST asynchronously, and the delays introduced by blocking until the protocol

has committed its delivery ordering could be significant. This issue is discussed further in [BCG91].

Our solution is based on the Isis replicated data update protocol described in [BJ89] and the

ABCAST protocol developed in [BJ87b, Sch88]. Associated with each view \( \textbf{view}_i(g) \) of a process
group \( g \) will be a \textit{token holder} process, \( \textbf{token}(g) \in \textbf{view}_i(g) \). We also assume that each message \( m \)
is uniquely identified by \( \textbf{uid}(m) \).

To ABCAST \( m \), a process holding the token uses CBCAST to transmit \( m \) in the normal manner.

If the sender is not holding the token, the ABCAST is done in stages:

1. The sender CBCAST's \( m \) but marks it as undeliverable.\(^1\) Processes other than the token

   holder (including the sender) that receive this message place \( m \) on the CBCAST delay queue
   in the usual manner, but do not remove \( m \) from the queue for delivery even after all messages
   that precede it causally have been delivered. It follows that a typical process may have some
   number of delayed ABCAST messages at the front of its CBCAST delay queue. This prevents
   the delivery of causally subsequent CBCAST messages, because the vector time is not updated

\(^1\)It might appear cheaper to forward such a message directly to the token holder. However, for a moderately large

messages such a solution would double the IO done by the token holder, creating a likely bottleneck, while reducing

the IO load on other destinations only to a minor degree.
until delivery occurs. On the other hand, a CBCAST that precedes or is concurrent with one of these undeliverable ABCAST messages will not be delayed.

2. The token holder treats incoming ABCAST messages as it would treat incoming CBCAST messages, delivering them in the normal manner. However, it also makes note of the uid of each such ABCAST.

3. After the process holding the token has delivered one or more ABCAST messages, it uses CBCAST to send a sets-order message giving a list of one or more messages, identified by uid, and ordered in the delivery order that arose in step 2. If desired, this CBCAST may be delayed so as to "batch" such transmissions, but it must be sent before (or piggybacked upon) any subsequent ABCAST or CBCAST by the token holder. If desired, a new token holder may be specified in this message.

4. On receipt of a sets-order message, a process places it on the CBCAST delay queue in the normal manner. Eventually, all the ABCAST messages referred to in the sets-order message will be received, and all the CBCAST messages that precede the sets-order will have been delivered (liveness of CBCAST).

Recall that \( \rightarrow \) places a partial order on the messages in the delay queue. Our protocol now re-orders concurrent ABCAST messages by placing them in the order given by the sets-order message, and marks them as deliverable.

5. Deliverable ABCAST messages may be delivered off the front of the queue.

Step 4 is the key one in the protocol. This step causes all participants to deliver ABCAST messages in the order that the token holder used. This order will be consistent with causality because the token holder itself treated these ABCAST messages as if they were CBCAST's.

The cost of doing an ABCAST depends on the locations where multicasts originate and the frequency with which the token is moved. If multicasts tend to originate at the same process repeatedly, then once the token is moved to that site, the cost is one CBCAST per ABCAST. If they originate randomly and the token is not moved, the cost is \(1 + 1/k\) CBCAST's per ABCAST, where we assume that one set-order message is sent for ordering purposes after \(k\) ABCAST's.

5.3 Multicast stability

The knowledge that a multicast has reached all its destinations will be useful below. Accordingly, we will say that a multicast \(m\) is \(k\)-stable if it is known that the multicast has been received at \(k\) destinations. When \(k = |\text{dests}(m)|\) we will say that \(m\) is (fully) stable.
Recall that our model assumes a reliable transport layer. Since messages can be lost in transport for a variety of reasons (buffer overflows, noise on communication lines, etc.), such a layer normally uses some sort of positive or negative acknowledgement scheme to confirm delivery. Our liveness assumption is strong enough to imply that any multicast \( m \) will eventually become stable, if the sender does not fail. It is thus reasonable to assume that this information is available to the process that sent a CBCAST message.

To propagate stability information among the members of a group, we will introduce the following convention. Each process \( p_i \) maintains a \textit{stability sequence number}, \( \text{stable}(p_i) \). This number will be the largest integer \( n \) such that all multicasts by sender \( p_i \) having \( VT(p_i)[i] \leq n \) are stable.

When sending a multicast \( m \), process \( p_i \) piggybacks its current value of \( \text{stable}(p_i) \); recipients make note of these incoming values. If \( \text{stable}(p_i) \) changes and \( p_i \) has no outgoing messages then, when necessary (see below), \( p_i \) can send a \textit{stability} message containing only \( \text{stable}(p_i) \).

5.4 VT compression

It is not always necessary to transmit the full vector timestamp on each message.

\textbf{Lemma 1:} Say that process \( p_i \) sends a multicast \( m \). Then \( VT(m) \) need only carry vector timestamp fields that have changed since the last multicast by \( p_i \).

\textbf{Proof:} Consider two multicasts \( m \) and \( m' \) such that \( m \rightarrow m' \). If \( p_i \) is the sender of \( m \) and \( p_j \) is the sender of \( m' \), then there are two cases. If \( i = j \) then \( VT(m)[i] < VT(m')[i] \), hence (step 2, Section 5.1) \( m' \) cannot be delivered until after \( m \) is delivered. Now, if \( i \neq j \) but \( m' \) carries the field for \( p_i \), then \( VT(m)[i] \leq VT(m')[i] \), and again, \( m' \) will be delayed under step 2 until after \( m \) is delivered. But, if this field is omitted, there must be some earlier message \( m'' \), also multicast by \( p_j \), that did carry the field. Then \( m \) will be delivered before \( m'' \) and, under the first case, \( m'' \) will be delivered before \( m' \). \( \Box \)

Compression may not always be advantageous: the data needed to indicate which fields have been transmitted may actually increase the size of the VT representation. However, in applications characterized by relatively localized, bursty communication, compression could substantially reduce the size of the timestamp. In fact, if a single process sends a series of messages, and receives no messages between the sends, then the VT timestamp on all messages but the first will just contain one field. Moreover, in this case, the value of the field can be inferred from the FIFO property of the channels, so such messages need not contain \textit{any} timestamp. We will make further use of this idea below.
5.5 Delivery atomicity and group membership changes

We now consider the implementation of atomicity and how group membership changes impact the above protocols. Such events raise several issues that are addressed in turn:

1. Virtually synchronous addressing.
2. Re-initializing VT timestamps.
3. Delivery atomicity when failures occur.
4. Handling delayed ABCAST messages when the token holder fails without sending a sets-order message.

Virtually synchronous addressing

To achieve virtually synchronous addressing when group membership changes while multicasts are active, we introduce the notion of flushing the communication in a process group. Consider a process group in view $i$. Say that view $i + 1$ now becomes defined. We can flush communication by having all the processes in $view_{i+1}$ send a message "flush i+1", to all other processes in this view. After sending such messages and before receiving such a flush message from all members of $view_{i+1}$ a process will accept and deliver messages but will not initiate new multicasts. Because communication is FIFO, if process $p$ has received a flush message from all processes in $view_{i+1}$, it will first have received all messages that were sent by members of $view_i$. In the absence of failures, this establishes that multicasts will be virtually synchronous in the sense of Section 4.

A disadvantage of this flush protocol is that it sends $n^2$ messages. Fortunately, the solution is readily modified into one that uses a linear number of messages. Still deferring the issue of failures, say that we designate one member of $view_{i+1}$, $p_c$, as the flush coordinator. Any deterministic rule can be used for this purpose. A process $p_i$ other than the coordinator flushes by first waiting until all multicasts that it sent have stabilized, and then sending a $view_{i+1}$ flush message to the coordinator. The coordinator, $p_c$, waits until flush messages have been received from all other members of $view_{i+1}$. It then multicasts its own flush message to the members of $view_{i+1}$ (it need not wait for its own multicasts to stabilize). Reception of the flush multicast from $p_c$ provides the same guarantees as did the original solution. However, the cost of the new protocol is much lower: $2n$ messages.
Reinitializing VT fields

After executing the flush protocol, all processes can reset the fields of VT to zero. This is clearly useful. Accordingly, we will now assume that the vector timestamp sent in a message, VT(m), includes the index i of the view viewi in which m was transmitted. A vector timestamp carrying an expired index may be ignored, since the flush protocol used to switch views will have forced the delivery of all multicasts that could have been pending in the prior view.

Delivery atomicity and virtual synchrony when failures occur

We now consider the case where some process fails during an execution. Failures introduce two basic problems:

1. A failure could disrupt the transmission of a multicast. Thus, if pj has received a multicast m message from pi, and has not learned of the stability of that multicast, some of the other destinations of m may not have received a copy.

2. We can no longer assume that all processes will respect the flush protocol, since the failure of a process pi could prevent it from sending flush messages for some view, even if that view reports pi as still operational. On the other hand, we also know that a view showing the failure of pi will eventually be received.

To solve the first problem, we will have all processes retain copies of the messages they receive. If pj detects the failure of pi, it will forward a copy of any unstable multicasts it has received from pi to other members of the group. All processes identify and reject duplicates. However, the second problem could now prevent the protocol from being respected, leaving the first problem unsolved, as illustrated in Figure 3. This shows that the two problems are closely related and have to be addressed in a coordinated way.

The solution to this atomicity and virtual synchrony problem is most readily understood in terms of the original n^2 message flush protocol. If we are running that protocol, it suffices to delay the installation of viewi+1 until, for some k ≥ 1, flush messages for viewi+k have been received from all processes in viewi+k ∩ viewi+1. Notice that a process may be running the flush protocol for viewi+k despite not yet having installed viewi+1.

More formally, the algorithm executed by a process p is as follows.
Figure 3: Group Flush Algorithm
1. On receiving \textit{view}_{i+k}, \( p \) increments a local variable \textit{inhibit\_sends}; while the counter remains greater than 0, new messages will not be initiated. \( p \) forwards a copy of any unstable message \( m \) that was sent in \textit{view}_j to all processes in \textit{view}_j \cap \textit{view}_{i+k} \footnote{In practice, it may be easier and faster to multicast all the unstable messages from \textit{view}_j to all processes in \textit{VIEW}_{i+k}. Processes not in \textit{view}_j \cap \textit{view}_{i+k} will discard this multicast.}, and then marks \( m \) as stable. Lastly, \( p \) sends "flush \( i+k \)" to each member of \textit{view}_{i+k}.

2. On receiving a copy of a message \( m \), \( p \) examines the index of the view in which \( m \) was sent. If \( p \) most recently installed \textit{view}_i and \( m \) was sent in \textit{view}(m) < \( i \), \( p \) ignores \( m \) as a duplicate. If \( m \) was sent in view \( i \), \( p \) applies the normal delivery rule, identifying and discarding duplicates in the obvious manner. If \( m \) was sent in view \textit{view}(m) > \( i \), \( p \) saves \( m \) until \textit{view}(m) has been installed.

3. On receiving "flush \( i+k \)" messages from all processes in \textit{view}_i \cap \textit{view}_{i+k}, \( p \) installs \textit{view}_{i+1} by delivering it to the application layer. It decrements the \textit{inhibit\_sends} counter, and, if the counter is now zero, permits new messages to be initiated.

   Any message \( m \) that was sent in \textit{view}_i and has not yet been delivered may be discarded. This can only occur if the sender of \( m \) has failed, and has previously received and delivered a message \( m' \) (\( m' \rightarrow m \)) that has now been lost. In such a situation, \( m \) is an orphan of a system execution that was "erased" by multiple failures.\footnote{Notice that in our model, even if process \( p \) accepts and delivers a message \( m \) under this protocol, the failure of \( p \) could lead to a situation in which \( m \) is not delivered to its other destinations. Our definition of delivery atomicity could be changed to exclude such executions, but atomicity could then only be achieved using a much more costly 2-phase delivery protocol.}

4. A message can be discarded as soon as it has been delivered locally and has become stable. Notice that a message becomes stable after having been forwarded at most once (in step 1).

5. A process \( p \) that sent a message \( m \) in view \( i \) will now consider \( m \) to have stabilized if delivery confirmation has been obtained from all processes in \textit{view}_i \cap \textit{view}_{i+k}, for any value of \( k \), or if \textit{VIEW}_{i+1} has been installed.

\textbf{Lemma 2:} The flush algorithm is safe and live.

\textbf{Proof:} A participant in the protocol, \( p \), delays installation of \textit{view}_{i+1} until it has received flush messages sent in view \( i + k \) (\( k \geq 1 \)) from all members of \textit{view}_{i+1} that are still operational in \textit{view}_{i+k}. These are the only processes that could have a copy of an unstable message sent in \textit{view}_i. This follows because messages from failed processes are discarded, and because messages are only forwarded to processes that were alive in the view in which they were originally sent. Because participants forward unstable messages before sending the flush, and channels are FIFO, \( p \) will have received all messages sent in \textit{view}_i before installing \textit{view}_{i+1}. It follows that the protocol is
safe. Liveness follows because the protocol to install view_i can only be delayed by the failure of a member of view_{i+1}. Since view_{i+1} has a finite number of members, any delay will be finite.

This protocol can be modified to send a linear number of messages in each view, using the same coordinator-based scheme as was proposed earlier. Instead of sending "flush i+k" messages to all processes in view_{i+1} \cap view_{i+k}, a process forwards copies of unstable messages to the coordinator, followed by a "flush i+k" message. (A message m originally sent in VIEW_i by a process p, which failed before m became stable, is regarded as unstable by a surviving process q \in VIEW_i until q installs VIEW_{i+1}.) The coordinator, in turn, forwards these messages back to the other processes in the view, and then, after receiving "flush i+k" messages from all the members in view_{i+k}, multicasts its own flush message. view_{i+k} can be installed when the flush message from the coordinator is received.

We now consider the case where failures occur during the execution of the above protocol. The coordinator should abort its protocol if it is waiting for a flush message from some process p, and view_{i+k+1} becomes defined, showing that p has failed. In this case, the protocol for view_{i+k+1} will subsume the one for view_{i+k}. Similarly, if the coordinator itself fails, a future flush run by some other coordinator will subsume the interrupted protocol. A successfully completed flush view will now permit installation of all prior views.

For liveness, it now becomes necessary to avoid infinitely delaying the installation of a view in the event of an extended process join/failure sequence. We therefore modify the protocol to have participants inform the coordinator of their current view. A coordinator that has gathered flush messages for view i+k from all the processes in view_{i+1} \cap view_{i+k} can send a view i+1 flush message to any process still in view i, even if it has not yet received all the flush messages needed to terminate the protocol for view i+k. With this change, the protocol is live.

As illustrated in Figure 3, this protocol converts an execution with non-atomic multicasts into one in which all multicasts are atomically delivered, at linear cost.

**ABCAST ordering when the token holder fails**

The atomicity mechanism of the preceding subsection requires a small modification of the ABCAST protocol. Consider an ABCAST that is sent in view_i, and for which the token holder fails before sending the sets-order message.

After completion of the flush protocol for view i, the ABCAST message will be on every delay queue, but not delivered anywhere. Moreover, any sets-order messages that were initiated before
the failure will have been delivered everywhere, hence the set of undelivered ABCAST messages is the same at all processes. These messages must be delivered before \( \text{view}_i \) can be installed.

Notice that the delay queue is partially ordered by \( \rightarrow \). We can solve our problem by ordering any concurrent ABCAST messages within this set using any well-known, deterministic rule. For example, they can be sorted by \( \text{uid} \). The resulting total order on ABCAST messages will be the same at all processes and consistent with causality.

6 Extensions to the basic protocol

Neither of the protocols in Section 5 is suitable for use in a setting with multiple process groups. We first introduce the modifications needed to extend CBCAST to a multi-group setting. We then briefly examine the problem of ABCAST in this setting.

The CBCAST solution we arrive at initially could be costly in systems with very large numbers of process groups or groups that change dynamically. This has not been a problem in the current ISIS system because current applications use comparatively few process groups, and processes tend to multicast for extended periods in the same group. However, these characteristics will not necessarily hold in future ISIS applications. Accordingly, the second part of the section explores additional extensions of the protocol that would permit its use in settings with very large numbers of very dynamic process groups. The resulting protocol is interesting because it exploits properties of what we call the communication structure of the system.

6.1 Extension of CBCAST to multiple groups

The first extension to the protocol is concerned with systems composed of multiple process groups. We will continue to assume that a given multicast is sent to a single group destination, but it may now be the case that a process belongs to several groups and is free to multicast in any of them.

Suppose that process \( p_i \) belongs to groups \( g_a \) and \( g_b \), and multicasts within both groups. Multicasts sent by \( p_i \) to \( g_a \) must be distinguished from those to \( g_b \). If not, a process belonging to \( g_b \) and not to \( g_a \) that receives a message with \( VT(m)[i] = k \) will have no way to determine how many of these \( k \) messages were actually sent to \( g_b \) and should, therefore, precede \( m \) causally.

This leads us to extend the single \( VT \) clock to multiple \( VT \) clocks. We will use the notation \( VT_a \).
to denote the logical clock associated with group $g_a$; $VT_a[i]$ thus counts\footnote{Clearly, if $p_i$ is not a member of $g_a$, then $VT_a[i] = 0$, allowing a sparse representation of the timestamp. For clarity, our figures will continue to depict each timestamp $VT_p$ as a vector of length $n$, with a special entry * for each process that is not a member of $g_a$.} multicasts by process $p_i$ to group $g_a$. The stability sequence number, $stable(p_i)$ should be similarly qualified: $stable_a(p_i)$. Processes maintain $VT$ clocks for each group in the system, and attach all the $VT$ clocks to every message that they multicast.

The next change is to step 2 of the protocol (Section 5.1). Suppose that process $p_j$ receives a message $m$ sent in group $g_a$ with sender $p_i$, and that $p_j$ also belongs to groups $\{g_1, ..., g_n\} \subseteq G_j$. Step 2 can be replaced by the following rule:

2' On reception of message $m$ from $p_i \neq p_j$, sent in $g_a$, process $p_j$ delays $m$ until

\begin{align*}
2.1' & \quad VT_a(m)[i] = VT_a(p_j)[i] + 1, \text{ and} \\
2.2' & \quad \forall k : (p_k \in g_a \land k \neq i) : VT_a(m)[k] \leq VT(p_j)[k], \text{ and} \\
2.3' & \quad \forall g : (g \in G_j) : VT_g(m) \leq VT_g(p_j).
\end{align*}

This is just the original protocol modified to iterate over the set of groups to which a process belongs. As in the original protocol, $p_j$ does not delay messages received from itself.

Figure 4 illustrates the application of this rule in an example with four processes identified as $p_1...p_4$. Processes $p_1$, $p_2$ and $p_3$ belong to group $G_1$, and processes $p_2$, $p_3$ and $p_4$ to group $G_2$. Notice that $m_2$ and $m_3$ are delayed at $p_3$, because it is a member of $G_1$ and must receive $m_1$ first. However, $m_2$ is not delayed at $p_4$, because $p_4$ is not a member of $G_1$. And $m_3$ is not delayed at $p_2$, because $p_2$ has already received $m_1$ and it was the sender of $m_2$.

The proof of Section 5 adapts without difficulty to this new situation; we omit the nearly identical argument. One can understand the modified protocol in intuitive terms. By ignoring the vector timestamps for certain groups in step 2.3', we are asserting that there is no need to be concerned that any undelivered message from these groups could causally precede $m$. But, the ignored entries correspond to groups to which $p_j$ does not belong. Since all communication is done within groups, these entries are irrelevant to $p_j$.

6.2 Multiple-group ABCAST

When run over this extended CBCAST protocol, our ABCAST solution will continue to provide a total, causal delivery ordering within any single process group. However, it will not order multicasts
to different groups even if those groups have members in common. In [BCG91] we examine the need for a global ABCAST ordering property and suggest that it may be of little practical importance, since the single-group protocol satisfies the requirements of all existing Isis applications that we know of. We have extended our ABCAST solution to a solution that provides a global total ordering; the resulting protocol, in any single group $g$, has a cost proportional to the number of other groups that overlap with $g$. Details can be found in [Ste91]. This extended, global-ordering ABCAST protocol could be implemented if the need arises.

6.3 Extended VT compression

In Section 5.4 we introduced a rule for compressing a vector timestamp before transmission. One might question the utility of such a technique within a single process group, especially if the group is likely to be small. In a multiple-group setting, the same technique might permit a process to omit entire vector timestamps from some of its multicasts.

Intuitively, the “latest” timestamp for a group $g_a$ need only be included on the first of any messages sent in some other group $g_b$. Further communication within $g_b$ need not include this timestamp, since all this communication will be causally after the message that contained the updated value of $VT_a$. More precisely, any process in $g_b$ that updates its value of $VT_a$ as a result of a message received from some process not in $g_b$ will include the updated value of $VT_a$ in the next message $m$ that it multicasts to $g_b$. The updated value need not be included in any subsequent messages multicast to $g_b$. Further details of this optimization, and a proof of correctness, can be found.
We can do even better. Recall the definition of multicast stability (Section 5.3). Say that a group is active for process $p$, if:

1. $p$ is the initiator of a multicast to $g$ that is not stable, or
2. $p$ has received an unstable multicast to $g$.

Activity is a local property; i.e. process $p$ can compute whether or not some group $g$ is active for it by examining local state. Moreover, $g$ may be active for $p$ at a time it is inactive for $q$ (in fact, by delaying the delivery of messages, a process may force a group to become inactive just by waiting until any active multicasts stabilize).

Now, we modify step 2' of the extended protocol as follows. A process $p$ which receives a multicast $m$ in group $g_a$ must delay $m$ until any multicasts $m'$ previously received in other groups $g_b$ ($b \neq a$) have been delivered locally. For example, say that process $p$ receives $m_1$ in group $g_1$, then $m_2$ in group $g_2$, and then $m_3$ in group $g_1$. Under the rule, $m_2$ must not be delivered until after $m_1$. Similarly, $m_3$ must not be delivered until after $m_2$. Since the basic protocol is live in any single group, no message will be delayed indefinitely under this modified rule.

Then, when sending messages (in any group), timestamps corresponding to inactive groups can be omitted from a message. The intuition here is that it is possible for a stable message to have reached its destinations, but still be blocked on some delivery queues. Our change ensures that such a message will be delivered before any subsequent messages received in other groups. Knowing that this will be the case, the vector timestamp can be omitted.

It is appealing to ask how effective timestamp compression will be in typical Isis applications. In particular, if the compression dramatically reduces the number of timestamps sent on a typical message, we will have arrived at the desired, low overhead, protocol. On the other hand, if compression is ineffective, measures may be needed to further reduce the number of vector timestamps transmitted. Unfortunately, we lack sufficient experience to answer this question experimentally. At this stage, any discussion must necessarily be speculative.

Recall from Section 2 that Isis applications are believed to exhibit communication locality. In our setting, locality would mean that a process that most recently received a message in group $g_i$ will probably send and receive several times in $g_i$ before sending or receiving in some other group. It would be surprising if distributed systems did not exhibit communication locality, since analogous properties are observed in almost all settings, reflecting the fact that most computations involve some form of looping [Den80]. Process-group based systems would also be expected to exhibit
locality for a second reason: when a distributed algorithm is executed in a group \( g \) there will often be a flurry of message exchange by participants. For example, if a process group were used to manage transactionally replicated data, an update transaction might multicast to request a lock, to issue each update, and to initiate the commit protocol. Such sequences of multicasts arise in many Isis algorithms.

The extended compression rule benefits from communication locality, since few vector timestamps would be transmitted during a burst of activity. Most groups are small, hence those timestamps that \( do \) need to be piggybacked on a message will be small. Moreover, in a system with a high degree of locality, each process group through which a vector timestamp passes will “delay” the timestamp briefly.

For example, suppose that a process that sends one multicast in group \( g_{1} \) will, on the average, send and receive a total of \( n \) multicasts in \( g_{1} \) before sending or receiving in some other group. Under our extended rule, only the first of these multicasts will carry vector timestamps for groups other than \( g_{1} \). Subsequent multicasts need carry no vector timestamps at all, since the sender’s timestamp can be deduced using the method of Section 5.4. Moreover, if the vector timestamp for a group \( g_{1} \) changes \( k \) times per second, members of an adjacent group \( g_{2} \) (that are not also members of \( g_{1} \)) will see a rate of change of \( k/n \). A group at distance \( d \) would see every \( n^{d} \)th value, giving a rate of \( k/n^{d} \) per second. Thus, the combination of compression and communication locality can substantially reduce the vector timestamp overhead on messages. In fact, if most messages are sent in bursts, the “average” multicast may not carry any timestamps at all!

6.4 Atomicity and group membership changes

The protocol for group flush and multicast atomicity need to be reconsidered in light of this multiple group extension.

Virtually synchronous addressing

Recall that virtually synchronous addressing is implemented using a group flush protocol. In the absence of failures, the protocol protocol of Section 5.5 continues to provide this property. Although it may now be the case that some messages arrive carrying “stale” vector timestamps corresponding to group views that are no longer installed, the convention of tagging the vector timestamp with the view index number for which it was generated permits the identification of these timestamps, which may be safely ignored: any messages to which they refer were delivered during an earlier view flush operation.
Atomicity and virtual synchrony when failures occur

When the flush algorithm of Section 5.5 is applied in a setting with multiple, overlapping groups, failures introduce problems not seen in single-group settings.

Consider two processes \( p_1, p_2 \) and two groups \( g_1, g_2 \), such that \( p_1 \) belongs to \( g_1 \) and \( p_2 \) to both \( g_1 \) and \( g_2 \), and suppose the following event sequence occurs:

1. \( p_1 \) multicasts \( m_1 \) to \( g_1 \) in view \( \text{view}(g_1) \).
2. \( p_2 \) receives and delivers \( m_1 \), while in view \( (\text{view}(g_1), \text{view}(g_2)) \).
3. \( p_2 \) multicasts \( m_2 \) to \( g_2 \), still in view \( (\text{view}(g_1), \text{view}(g_2)) \).
4. Both \( p_1 \) and \( p_2 \) fail, causing the installation of new views \( \text{view}(g_1)' \) and \( \text{view}(g_2)' \).

Now, consider a process \( q \) belonging to both \( g_1 \) and \( g_2 \). This process will be a destination of both \( m_1 \) and \( m_2 \). If \( p_2 \) was the only process to have received \( m_1 \) before \( p_1 \) failed, \( m_1 \) will be lost due to the failure; \( q \) would then install \( \text{view}(g_1)' \) without delivering \( m_1 \). Further, suppose that \( m_2 \) has been received by another process \( q' \), belonging to \( g_2 \) but not \( g_1 \). If \( q' \) remains operational, \( q \) will receive \( m_2 \) during the flush protocol for \( \text{view}(g_2)' \). This creates a problem:

1. If \( q \) delivers \( m_2 \) before installing \( \text{view}(g_1)' \), causality will be violated, because \( m_1 \) was not delivered first.
2. If \( m_2 \) is not delivered by \( q \), atomicity will be violated, because \( m_2 \) was delivered at a process \( q' \) that remained operational.

Even worse, \( q \) may not be able to detect any problem in the first case. Here, although \( m_2 \) will carry a vector timestamp reflecting the transmission of \( m_1 \), the timestamp will be ignored as "stale." In general, \( q \) will only recognize a problem if \( m_2 \) is received before the flush of \( g_1 \) has completed.

There are several ways to solve this problem. Since the problem does not occur when a process communicates only within a single group, our basic approach will be to intervene when a process begins communicating with another group, delaying communication in group \( g_2 \) until causally prior messages to group \( g_1 \) are no longer at risk of loss. Any of the following rules would have this effect:

- One can build a \( k \)-resilient protocol that operates by delaying communication outside a group until all causally previous messages are \( k \)-stable; as \( k \) approaches \( n \), this becomes a conservative but completely safe approach. Here, the sender of a message may have to wait before being permitted to transmit it.
• Rely on the underlying message transport protocol to deliver the message to all destinations despite failures. For example, a token ring or ethernet might implement reliable multicast transport directly at the link level, so that messages are never lost at this stage.

• Construct a special-case message logging server which is always guaranteed to have a copy of any messages that have not yet been delivered to all their destinations. Such a facility would resemble the associative message store in the original Isis system [BJ87b] and also recalls work by Presotto and Powell [Pow83].

Our implementation uses the first of these schemes, and as initially structured, operates completely safely (i.e. with \( k = n \)). That is, a process \( p_i \) that has sent or received multicasts in group \( g_1 \) will delay initiation of a multicast in group \( g_2 \) until the \( g_1 \) multicasts are all fully stable. Our future system will be somewhat more flexible, allowing the process that creates a group to specify a value of \( k \) for that group; such an approach would have a performance benefit. We note that standard practice in industry is to consider a system fault-tolerant if it can tolerate a single failure, i.e. \( k = 1 \).

Barry Gleeson\(^5\) has made the following observation. Delays associated with multicast stability for reasons of atomicity represent the most likely source of delay in our system. However, with the following sort of specialized multicast transport protocol, this delay can be completely eliminated. Consider an application in which the sender of a message is often co-resident with one of the recipients, that is, on the same physical computer, and in which the network interface is reliable (lossless) and sequenced (sends messages in order, even when the destination sets differ). This implies, for example, that if a message \( m \) is received by a process \( p \), any message \( m' \) transmitted from the same source prior to \( m \) will have already been received at all its destinations.

On the resulting system, it would never be necessary to delay messages to a local destination: any action taken by a local recipient and causally related to the received message would only reach external nodes after stability of the prior multicast. Moreover, a remote destination would never need to delay a message because it could safely assume that the other destinations have also received it. In this setting, no message need ever be delayed due to atomicity considerations! Our protocols could then perform particularly well, since they would tend to "pipeline" multicasts between nodes, while never delaying intra-node communication at all.

6.5 Use of communication structure

Until the present, we have associated with each message a vector time or vector times having a total size that could be linear in the number of processes and groups comprising the application.

\(^5\)Barry Gleeson is with the UNISYS Corporation, San Jose, California.
On the one hand, we have argued that in many systems compression could drastically reduce size. Moreover, similar constraints arise in many published CBCAST protocols. However, one can still imagine executions in which vector sizes would grow to dominate message sizes. A substantial reduction in the number of vector timestamps that each process must maintain and transmit is possible in the case of certain communication patterns, which are defined precisely below. Even if communication does not always follow these patterns, our new solution can form the basis of other slightly more costly solutions which are also described below.

Our approach will be to construct an abstract description of the group overlap structure for the system. This structure will not be physically maintained in any implementation, but will be used to reason about communication properties of the system as it executes. Initially, we will assume that group membership is “frozen” and that the communication structure of the system is static. Later, in Sections 6.6 and 6.6, we will extend these to systems with dynamically changing communication structure. For clarity, we will present our algorithms and proofs in a setting where timestamp compression rules are not in use; the algorithms can be shown to work when timestamp compression is in use, but the proofs are more complicated.

Define the communication structure of a system to be an undirected graph $CG = (G, E)$ where the nodes, $G$, correspond to process groups and edge $(g_1, g_2)$ belongs to $E$ iff there exists a process $p$ belonging to both $g_1$ and $g_2$. If the graph so obtained has no biconnected component containing more than $k$ nodes, we will say that the communication structure of the system is $k$-bounded. In a $k$-bounded communication structure, the length of the largest simple cycle is $k$. A 0-bounded communication structure is a tree (we neglect the uninteresting case of a forest). Clearly, such a communication structure is acyclic.

Notice that causal communication cycles can arise even if $CG$ is acyclic. For example, in figure 4, messages $m_1$, $m_2$, $m_3$ and $m_4$ form a causal cycle spanning both $g_1$ and $g_2$. However, the acyclic structure restricts such communication cycles in a useful way. Below, we demonstrate that it is unnecessary to transport all vector timestamps on each message in the $k$-bounded case. If a given group is in a biconnected component of size $k$, processes in this group need only to maintain and transmit timestamps for other groups in this biconnected component. We can also show that they need to maintain at least these timestamps. As a consequence, if the communication structure is acyclic, processes need only maintain the timestamps for the groups to which they belong.

We proceed to the proof of our main result in stages. First we address the special case of an acyclic communication structure, and show that if a system has an acyclic communication structure, each

---

6Two vertices are in the same biconnected component of a graph if there is still a path between them after any other vertex has been removed.

7The nodes of a simple cycle (other than the starting node) are distinct; a complex cycle may contain arbitrary repeated nodes.
process in the system only maintains and multicasts the VT timestamps of groups to which it belongs. Notice that this bounds the overhead on a message in proportion to the size and number of groups to which a process belongs.

We will wish to show that if message \( m_1 \) is sent (causally) before message \( m_k \), then \( m_1 \) will be delivered before \( m_k \) at all overlapping sites. Consider the chain of messages below.

\[
\begin{align*}
  m_1 & \Rightarrow m_2 & \Rightarrow m_3 & \Rightarrow \cdots & \Rightarrow m_{k-1} & \Rightarrow m_k & \Rightarrow m_{k+1} \\
  p_1 & \Rightarrow p_2 & \Rightarrow p_3 & \Rightarrow \cdots & \Rightarrow p_{k-1} & \Rightarrow p_k & \Rightarrow p_{k+1} \\
  g_1 & \Rightarrow g_2 & \Rightarrow g_3 & \Rightarrow \cdots & \Rightarrow g_{k-1} & \Rightarrow g_k
\end{align*}
\]

This schema signifies that process \( p_1 \) multicasts message \( m_1 \) to group \( g_1 \), that process \( p_2 \) first receives message \( m_1 \) as a member of group \( g_1 \) and then multicasts \( m_2 \) to \( g_2 \), and so forth. In general, \( g_i \) may be the same as \( g_j \) for \( i \neq j \) and \( p_i \) and \( p_j \) may be the same even for \( i \neq j \) (in other words, the processes \( p_i \) and the groups \( g_i \) are not necessarily all different). Let the term message chain denote such a sequence of messages, and let the notation \( \text{m}_i \vdash \text{m}_j \) mean that \( p_j \) transmits \( m_j \) using a timestamp \( \text{VT}(m_j) \) that directly reflects the transmission of \( m_i \). For example, say that \( m_i \) was the \( k \)'th message transmitted by process \( p_i \) in group \( g_a \). So \( \text{m}_i \vdash \text{m}_j \) iff \( \text{VT}_a(p_j)[i] \geq k \) and consequently \( \text{VT}_a(m_j)[i] \geq k \). Our proof will show that if \( \text{m}_i \vdash \text{m}_j \) and the destinations of \( m_i \) and \( m_j \) overlap, then \( \text{m}_i \vdash \text{m}_j \), where \( p_j \) is the sender of \( m_j \). Consequently, \( m_i \) will be delivered before \( m_j \) at any overlapping destinations.

We now note some simple facts about this message chain that we will use in the proof. Recall that a multicast to a group \( g_a \) can only be performed by a process \( p_i \) belonging to \( g_a \). Also, since the communication structure is acyclic, processes can be members of at most two groups. Since \( m_k \) and \( m_1 \) have overlapping destinations, and \( p_2 \), the destination of \( m_1 \), is a member of \( g_1 \) and of \( g_2 \), then \( g_k \), the destination of the final broadcast, is either \( g_1 \) or \( g_2 \). Since \( CG \) is acyclic, the message chain \( m_1 \ldots m_k \) simply traverses part of a tree reversing itself at one or more distinguished groups. We will denote such a group \( g_r \). Although causality information is lost as a message chain traverses the tree, we will show that when the chain reverses itself at some group \( g_r \), the relevant information will be "recovered" on the way back.

**Lemma 3:** If a system has an acyclic communication structure, each process in the system only maintains and multicasts the VT timestamps of groups to which it belongs.

**Proof:** The proof is by induction on \( l \), the length of the message chain \( m_1 \ldots m_k \). Recall that we must show that if \( m_1 \) and \( m_k \) have overlapping destinations, they will be delivered in causal order at all such destinations, i.e, \( m_1 \) will be delivered before \( m_k \).
**Base case.** \( l = 2.\) Here, causal delivery is trivially achieved, since \( p_k \equiv p_2 \) must be a member of \( g_1 \) and \( m_k \) will be transmitted with \( g_1 \)'s timestamp. It will therefore be delivered correctly at any overlapping destinations.

**Inductive step.** Suppose that our algorithm delivers all pairs of causally related messages correctly if there is a message chain between them of length \( l < k.\) We show that causality is not violated for message chains where \( l = k.\)

Consider a point in the causal chain where it reverses itself. We represent this by \( m_{r-1} \rightarrow m_r \rightarrow m_{r+1}, \) where \( m_{r-1} \) and \( m_{r+1} \) are sent in \( g_{r-1} \equiv g_{r+1} \) by \( p_r \) and \( p_{r+1} \) respectively, and \( m_r \) and \( m_{r+1} \) are sent in \( g_r \) by \( p_r \) and \( p_{r+1} \). Note that \( p_r \) and \( p_{r+1} \) are members of both groups. This is illustrated in Figure 5. Now, \( m_r \) will not be delivered at \( p_{r+1} \) until \( m_r \) has been delivered there, since they are both broadcast in \( g_r.\) We now have \( m_{r-1} \nsubseteq m_r \nsubseteq m_{r+1}.\) We have now established a message chain between \( m_1 \) and \( m_k \) where \( l < k.\) So, by the induction hypothesis, \( m_1 \) will be delivered before \( m_k \) at any overlapping destinations, which is what we set out to prove. \( \square \)

**Theorem 1:** Each process \( p_i \) in a system needs only to maintain and multicast the VT timestamps of groups in the biconnected components of \( CG \) to which \( p_i \) belongs.

**Proof:** As with Lemma 3, our proof will focus on the message chain that established a causal link between the sending of two messages with overlapping destinations. This sequence may contain simple cycles of length up to \( k,\) where \( k \) is the size of the largest biconnected component of \( CG.\) Consider the simple cycle illustrated below, contained in some arbitrary message chain.
Now, since \( p_1, p_2 \) and \( p_3 \) are all in groups in a simple cycle of \( CG \), all the groups are in the same biconnected component of \( CG \), and all processes on the message chain will maintain and transmit the timestamps of all the groups. In particular, when \( m_1 \) arrives at \( p_3 \), it will carry a copy of \( VT_{g_1} \), indicating that \( m_1 \) was sent. This means that \( m_1 \) will not be delivered at \( p_3 \) until \( m_1 \) has been delivered there. So \( m_{l+1} \) will not be transmitted by \( p_3 \) until \( m_1 \) has been delivered there. Thus \( m_1 \rightarrow m_{l+1} \). We may repeat this process for each simple cycle of length greater than 2 in the causal chain, reducing it to a chain within one group. We now apply Lemma 3, completing the proof. \( \square \)

Theorem 1 shows us what timestamps are sufficient to assure correct delivery of messages. Are all these timestamps in fact necessary? It turns out that the answer is yes. It is easy to show that if a process that is a member of a group within a biconnected component of \( CG \) does not maintain a \( VT \) timestamp for some other group in \( CG \), causality may be violated. We therefore state without formal proof:

**Theorem 2:** If a system uses the VT protocol to maintain causality, it is both necessary and sufficient for a process \( p_i \) to maintain and transmit those VT timestamps corresponding to groups in the biconnected component of \( CG \) to which \( p_i \) belongs.

### 6.6 Extensions to arbitrary, dynamic communication structures

The previous section assumed that the communication graph was known statically. Operationally, this would correspond to a system in which, once established, process group membership never changed. Any realistic application is likely to be more dynamic, making it hard to manage information concerning the biconnected components of \( CG \). Moreover, any real distributed system will probably contain a mixture of subsystems, some having a regular communication structure, and some not.

Consider the multiple-group examples raised in the introduction. A scientific computation using groups for nearest neighbor communication will have an regular communication structure. The structure is known in advance and is a property of the algorithm, and it would be desirable to exploit this to reduce overhead on messages. Lemma 3 and Theorem 1 are ideally suited for this purpose. (We describe the system call interface used to communicate this information in Section 7).

This might or might not be true for the network information service. The designer of such a service has a choice between controlling the location of replication domains, or permitting data to migrate.
in an uncontrolled manner, creating fairly random domains. However, such a service will want
to balance processor loads and storage utilization, which might be hard in the latter approach.
Thus, the designer might well prefer to “tile” the network with process groups in a regular way,
which could then be exploited using our above results - again, presuming a suitable interface for
communicating this information.

On the other hand, a system built of abstract objects will almost certainly have an arbitrary
communication structure that changes continuously as applications are started and terminate.
Here, the communication structure would be impossible to represent; indeed, it may well change
faster than information about it can be propagated within the system. The best a process could
possibly do is to reason about local properties of the structure.

We now develop some simple results that enable processes to maintain only timestamps for groups
to which they actually belong, and yet to operate safely in dynamically changing communication
graphs that may contain cycles. Below, we will assume that processes may join and leave groups
dynamically, and may leave a group for reasons other than failure (in existing Isis this is possible,
but uncommon). This results in a highly dynamic environment. Nonetheless, a process might be
able to infer that a group to which it belongs is not present in any cycle. This would follow, for
example, if the group is adjacent to at most one other group. Such information can be obtained by
an exchange of adjacency information when a process joins a group, and subsequently multicasting
updated information in $g_1$ each time a current member joins some other group $g_2$. Further, say
that group $g_2$ is adjacent to groups $g_1$ and $g_3$, but that $g_1$ is adjacent to no other group. Then $g_2$
and eventually $g_3$ may come to learn that there is no cycle present.

**Conservative solution**

Our first solution is called the **conservative protocol**, and uses multicast stability (Section 5.3). The
idea will be to restrict the initiation of new multicasts so that a message $m$ can only be sent in a
group $g$ when it is known that any causally prior messages $m'$ will be delivered first, if $m$ and $m'$
share destinations.

The **conservative multicast rule** states that a process $p$ may multicast to group $g_1$ iff $g_1$ is the only
active group for process $p$ or $p$ has no active groups (the notion of an active group was defined in
Section 6.3). If $p$ attempts to multicast when this rule is not satisfied, it is simply delayed. During
this delay, incoming messages are not delivered. This means that all groups will eventually become
inactive, and the rule above will eventually be satisfied. At this point, the message is sent. It is now
immediate from the extended compression rule of Section 6.3 that when a message $m$ is multicast
in group $g$, only the sender’s timestamp for group $g$ need be transmitted.
The conservative rule imposes a delay only when two causally successive messages are sent to different groups. Thus, the rule would be inexpensive in systems with a high degree of locality. On the other hand, the overhead imposed would be substantial if processes multicast to several different groups in quick succession.

**Multicast epochs**

We now introduce an approach capable of overcoming the delays associated with the conservative rule, but at the cost of additional group flushes. We will develop this approach by first assuming that a process leaves a group only because of a failure, and then extending the result to cover the case of a process that leaves a group but remains active (as will be seen below, this can create a form of phantom cycle).

Assume that $CG$ contains cycles, but that some mechanism has been used to select a subset of edges $X$ such that $CG' = (G, E - X)$ is known to be acyclic. We extend our solution to use the acyclic protocol proved by lemma 2 for most communication within groups. If there is some edge $(g, g') \in X$, we will say that one of the two groups, say $g$, must be designated as an excluded group. In this case, all multicasts to or from $g$ will be done using the protocol described below.

Keeping track of excluded groups could be difficult; however it is easy to make pessimistic estimates (and we will derive a protocol that works correctly with such pessimistic estimates). For example, in Isis, a process $p$ might assume that it is in an excluded group if there is more than one other neighboring group. This is a safe assumption; any group in a cycle in $CG$ will certainly have two neighboring groups. This subsection develops solutions for arbitrary communication structures, assuming that some method such as the previous is used to safely identify excluded groups.

We will define a notion of multicast epoch, to be associated with messages such that if for two messages $m_1$ and $m_2$, $epoch(m_1) < epoch(m_2)$, then $m_1$ will always be delivered before $m_2$. In situations where a causal ordering problem could arise, our solution will increment the epoch counter. Specifically, each process $p$ maintains a local variable, $epoch_p$. When process $p$ initiates a multicast, it increments its epoch variable if the condition given below holds. It then piggybacks $epoch_p$ on the outgoing message. On reception of a message $m$, if $epoch_p < epoch(m)$, then $p$ will initiate the flush protocol for all groups to which it belongs, by sending a message "**start flush**" to the other group members. Reception of this message triggers execution of the flush protocol of Section 5.5, just as for a new group view (because our implementation clears vector timestamps as part of the flush, Isis numbers views using a major and minor view number: the major number is incremented for each new view, and the minor one for each flush done within the same view). On completing
the flush protocol, all group members set the value of their epoch variables to the maximum of the ones held by group members.

When will a process increment its epoch variable? Say that m is about to be multicast by p to g. We say that p is not safe in g if:

- The last message p received was from some other group g', and
- Either g or g' is an excluded group.

Our protocol rule is simple; on sending, if process p is not safe in group g, p will increment epoch_p before multicasting a message m to g. In this case, m will carry the epoch value but need not carry any vector timestamps. Otherwise, p will just increment its VT timestamp in the usual manner, and then piggyback onto m the epoch variable and timestamps for any (active) groups to which it belongs. A message is delivered when it is deliverable according to both the above flush rule and the VT delivery rule.

Notice that the flushing overhead of the modified protocol is incurred only when epoch values actually change, which is to say only on communication within two different groups in immediate succession, where one of the groups is excluded. That is, if process p executes for a period of time using the VT protocol and receives only messages that leave epoch_p unchanged, p will not initiate a flush. However, when an epoch variable is incremented the result could be a cascade of group flushes. Epoch variables will stabilize at the maximum existing value and flushing will then cease.

Theorem 3: The VT protocol extended to implement multicast epochs will deliver messages causally within arbitrary communication structures.

Proof: Consider an arbitrary message chain where the first and last messages have overlapping destinations. For example, in the chain shown below, p_{k+1} might be a member of both g_i and g_k and hence a destination of both m_1 and m_k. Without loss of generality, we will assume that g_1...g_k are distinct. We wish to show that the last message will be delivered after the first at all such destinations.

An interesting variation on this scheme would involve substituting synchronized realtime clocks for the epoch variable; a message would then be delayed until the epoch variable for a recipient advanced beyond some minimum value. Readers familiar with the Δ-T real-time protocols of [CASD86] will note the similarity between that protocol and such a modification of ours. In fact, clock synchronization (on which the Δ-T scheme is based) is normally done using periodic multicasts [ST87,Lam78], much like our flushing scheme. The development of a Δ-T based epoch protocol clearly warrants future study.
If none of \( g_1 \ldots g_k \) is an excluded group, then, by Lemma 3, \( m_1 \) will be delivered before \( m_k \) at \( p_{k+1} \). Now, let \( g_i \) be the first excluded group in the above message chain. If \( g_i \) is excluded, then \( p_i \) will increment its epoch variable before sending \( m_i \). As epoch numbers can never decrease along a causal chain, we will have \( \text{epoch}(m_1) < \text{epoch}(m_k) \), and a flush protocol will have run in any groups to which a destination of \( m_k \) belongs, before \( m_k \) can be delivered. \( m_1 \) was sent before the flush, and hence will be delivered by \( p_{k+1} \) before it delivers \( m_k \). □

Recall that we have been assuming that a process only leaves a group because of failure. Now, without changing the definition of the communication graph, say that processes can also leave groups for other reasons, remaining active and possibly joining other groups. Earlier, it was suggested that a process might observe that the (single) group to which it belongs is adjacent to just one other group, and conclude that it cannot be part of a cycle. In this class of applications, this rule may fail. The implication is that a process that should have incremented its epoch variable may neglect to do so, leading to a violation of the causal delivery ordering.

To see how a problem could arise, suppose that a process \( p \) belongs to group \( g_1 \), then leaves \( g_1 \) and joins \( g_2 \). If there was no period during which \( p \) belonged to both \( g_1 \) and \( g_2 \), \( p \) would use the acyclic VT protocol for all communication in both \( g_1 \) and \( g_2 \). Yet, it is clear that \( p \) represents a path by which messages sent in \( g_2 \) could be causally dependent upon messages \( p \) received in \( g_1 \), leading to a cyclic message chain that traverses \( g_1 \) and \( g_2 \). This creates a race condition under which violations of the causal delivery ordering could result.

This problem can be overcome in the following manner. Each process \( p_i \) simply increments its epoch counter after it leaves a group \( g_a \) and before it sends any subsequent messages. This will ensure that any message sent by \( p_i \) after it leaves \( g_a \) will be delivered subsequent to any message whose effect \( p_i \) had observed, directly or indirectly, before it left \( g_a \).

7 Applying the protocols to ISIS

This section briefly discusses some pragmatic considerations that arise when implementing the protocols for use in Isis.
Peer Groups | \(k, k \leq n\) (depending on degree of compression)  
| \(n\) is the number of group members.

Client/Server Groups | \(k, k \leq s + c\)  
| \(s\) is the number of servers and \(c\) is the number of clients.

Diffusion Groups | \(k, k \leq b\)  
| \(b\) is the number of members broadcasting into the group.

Hierarchical Groups | \(k, k \leq n\)  
| \(n\) is the size of a subgroup.

Table 1: Timestamp sizes resulting from different process group styles

7.1 Optimizations for Client/Server Groups

Up to now, our discussion has focused on communication in peer groups. In Isis client/server settings, a set of servers forms one group, and each client of the service they are providing forms an additional group containing that client and the server set (see figure 1). Theorem 2 appears to state that in this case, each group containing one of the clients needs to maintain the timestamps of every other such client group - a total timestamp size of \(O(s \times c)\), where \(s\) is the number of servers and \(c\) is the number of clients. Fortunately, since clients will not be communicating with each other except through the servers, and the servers form a peer subgroup of size \(s\) that receives all of these multicasts, an optimization can be applied to reduce the entire timestamp size to \(O(s + c)\). This optimization essentially collapses all the client groups into one large group; it is fully described in [Ste91]. A modified version of the conservative delay rule can be used to reduce any timestamps transmitted outside of the group to size \(s\). Finally, since timestamps associated with inactive groups can be omitted, and most clients of a large group will presumably be inactive, even the internal timestamps can be reduced in size, to length \(O(s + k)\) for some small \(k\).\(^9\)

Our protocols also work well with the other styles of process group usage, as summarized in Table 1. In diffusion groups, one vector timestamp is needed for the whole group. In the case of diffusion groups, the number of entries in the timestamp can be optimized: entries are only needed for the server processes, since these are the only ones that initiate multicasts. Hierarchical groups fall naturally into our general approach: since a process normally interacts with a single representative subgroup, the length of the vector timestamp seen is normally determined by the size of that subgroup. Further, a hierarchical group manager might have the freedom to create subgroups to be explicitly acyclic.

Some Isis applications form large process groups but would benefit from the ability to multicast to

---

9For example, by implementing the compression scheme or by simply having a client of a group drop out of it after some period of inactivity.

37
subsets of the total membership. For example, a stock quote service might send an IBM quote to only those brokers actually trading IBM stock. Such an ability can be emulated by forming groups approximating the multicast destination sets and discarding unwanted messages. Alternatively, our protocol can be extended into one directly supporting subset multicast. The basic idea is to move to a large timestamp, representing the number of times each possible sender has sent to each possible recipient. The resulting array would be sparse and repetitious, and hence could be substantially compressed. At present, we favor the former scheme, as it requires no changes to our basic protocol.

7.2 Point-to-point messages

Early in the paper, we asserted that asynchronous CBCAST is the dominant protocol used in Isis. Point-to-point messages, arising from replies to multicast requests and RPC interactions, are also common. In both cases, causal delivery is desired. Our implementation supports the transmission of point-to-point messages with causal delivery guarantees. This is done using an RPC scheme, in which the sender is inhibited from starting new multicasts until delivery of the point-to-point message is acknowledged. The vector time of the sender is included on the message, but the sender’s vector timestamp is not incremented prior to transmission. Point-to-point messages are thus treated as events internal to the processes involved.

The argument in favor of this method is that a single point-to-point RPC is fast and the cost is unaffected by the size of the system. Although one can devise more complex methods that eliminate the period of inhibited multicasting, problems of fault-tolerance render them less desirable.

7.3 System interface issues

One question raised by our protocols concerns the mechanism by which the system would actually be informed about special application structure, such as an acyclic communication structure. This is not an issue in the current Isis implementation, which uses the conservative rule, excluding groups adjacent to more than one neighboring group. In the current system, the only problem is to detect clients of a group, and as noted earlier, these declare themselves through the pg_client interface.

In the future, we expect to address these issues through a new system construct, the causality domain [BCG91]. A causality domain is a set of groups within which causality is enforced. Each group is created in a domain and subsequently remains in it. Causality is not enforced between domains, but a flush primitive can be provided, which will block a process until all causally prior
messages have been delivered. We are designing an interface by which parameters such as the message stability constant $k$ or assertions about the communication structure can be asserted on a per-domain basis. In the case of a domain declared to have an acyclic communication structure, a routine $\text{pg\_exclude}$ may be used to designate excluded groups.

An application such as the physics simulation described earlier could set up its group structure as a separate causality domain, declaring it to be acyclic and excluding enough groups to ensure that cycles will be broken. By invoking $\text{flush}$ before switching from domain to domain, causal safety would be achieved on both intra- and inter-domain operations.

### 7.4 Group view management

The current Isis implementation retains the prior Isis group view management server. Looking to the future, however, it will be possible to use our new protocol in a stand-alone fashion. Membership changes (adding or dropping a member from a group) can be implemented using the CBCAST protocol, including the new member as a destination in the former case. Clearly, this requires a form of mutual exclusion, which is obtained by having a distinguished group member initiate any such membership changes. Reception of this CBCAST message would trigger a view flush protocol: approximately $3n$ messages are thus needed to add one member to a group of size $n$. The addition or deletion of a client in a client-server or diffusion group is cheaper: a multicast and flush are still needed, but since clients don’t need to know about other clients, it can be confined to the server subgroup.

A source of failure information is needed in this scheme. In [RB91] we discuss the asynchronous failure detection problem and present an optimal algorithm.

Such a redesigned system would realize a long-term goal of our effort. The current Isis architecture is difficult to scale to very large LAN settings because of its reliance on a central protocol server. Although this server need not reside directly on every node, it introduces a bottleneck that limits the scalability of the architecture to networks with at most a few hundred workstations. By reimplementing the group view mechanism in terms of our new multicast protocol and separating failure detection into a free-standing module, this limit to scalability can be eliminated.

---

10 There are a number of possible implementations of $\text{flush}$, but for brevity we defer discussion of this issue to a future paper.

11 As in conventional distributed computing environments, this approach assumes that groups would be registered with some sort of group location service. Initial connection to a group would be via a forwarded request, causing the caller to be added as a new member or client. Subsequent to this, group operations could be performed directly.
8 Performance and transport protocol selection

We implemented the new CBCAST and ABCAST protocols within the Isis Toolkit, version 2.1 (publically available since Oct. 1990). The implementation is somewhat less ambitious than what we plan for the future Isis system. Nonetheless, the measured performance was encouraging.

<table>
<thead>
<tr>
<th>Group Size</th>
<th>Causal multicast with replies (ms)</th>
<th>Asynchronous msg/sec</th>
<th>kb/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>1</td>
<td>1.42</td>
<td>1.46</td>
<td>1.44</td>
</tr>
<tr>
<td>2</td>
<td>7.36</td>
<td>8.79</td>
<td>18.3</td>
</tr>
<tr>
<td>3</td>
<td>9.66</td>
<td>11.1</td>
<td>26.6</td>
</tr>
<tr>
<td>4</td>
<td>12.0</td>
<td>15.1</td>
<td>35.4</td>
</tr>
<tr>
<td>5</td>
<td>16.4</td>
<td>17.8</td>
<td>44.3</td>
</tr>
<tr>
<td>6</td>
<td>17.8</td>
<td>21.1</td>
<td>55.4</td>
</tr>
<tr>
<td>7</td>
<td>21.5</td>
<td>26.2</td>
<td>60.4</td>
</tr>
<tr>
<td>8</td>
<td>25.2</td>
<td>30.1</td>
<td>66.6</td>
</tr>
</tbody>
</table>

Table 2: Multicast performance figures. S: null packets M: 1K packets; L: 7K packets. All figures measured on Sun 4/60's running SUNOS 4.0.3.

Table 2 shows that the cost of transmitting to a group grows roughly linearly with the size of the group. The data was generated using the new CBCAST protocol within a single group. The figures are comparable to what can be achieved using the vendor-supplied remote procedure mechanism for the machines with which we worked: for example, a remote procedure call using a 1kb message through our facility (with a null reply) had a round-trip latency of 8.79ms. The figure for SUN RPC is comparable. Similarly, the throughput figures compare quite well with traditional streaming protocols such as TCP (we obtained 673 null messages per second, or 600kb of data per second). A graph of this data appears as Figure 6.

In order to understand how CPU time was expended by our protocol, we profiled the protocol within a single processor and between a pair of processes using RPC over a 10Mbit ethernet. Both request and reply were sent as single-destination CBCAST's. (The choice of RPC may seem odd, in light of our emphasis on asynchronous multicast, but this is the obvious way to measure round-trip delays). We then computed the costs attributable to different parts of the system. Table 3 shows a profile for a null RPC sent by a thread to another entry point within its own address space. This involves creating a new thread to handle each delivered message but no communication outside of
Table 3: Null causal multicast to another local thread w/reply

<table>
<thead>
<tr>
<th>Operation</th>
<th>Total Cost</th>
<th>Causal Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>send</td>
<td>406 uS</td>
<td>86 uS</td>
</tr>
<tr>
<td>receive</td>
<td>394 uS</td>
<td>31 uS</td>
</tr>
<tr>
<td>reply</td>
<td>406 uS</td>
<td>86 uS</td>
</tr>
<tr>
<td>receive</td>
<td>224 uS</td>
<td>31 uS</td>
</tr>
<tr>
<td>Total</td>
<td>1.43 mS</td>
<td>234 uS</td>
</tr>
<tr>
<td>(Measured)</td>
<td>1.43 mS</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: 1k causal multicast to a remote process w/reply

<table>
<thead>
<tr>
<th>Causal multicast</th>
<th>.23 mS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest of Isis</td>
<td>1.21 mS</td>
</tr>
<tr>
<td>Transport</td>
<td>3.6 mS</td>
</tr>
<tr>
<td>System Calls</td>
<td>1.59 mS</td>
</tr>
<tr>
<td>Wire Time</td>
<td>1.8 mS</td>
</tr>
<tr>
<td>Total</td>
<td>8.42 mS</td>
</tr>
<tr>
<td>(Measured)</td>
<td>8.79 mS</td>
</tr>
</tbody>
</table>

The address space of the test program. The first column shows total (measured) costs, the second shows costs attributed to causality, which we obtained by comparing the costs of sending causally ordered and FIFO messages.

Table 4 shows the costs on a layer-by-layer basis for the 8.79ms RPC to a remote process. The costs are broken down into the time spent in the protocol implementation (taken from the null multicast table, above), the transport costs, costs spent in system calls, and the time on the wire for a 1k message with its Isis-supplied header and a null reply. The header size used was approximately 400 bytes in each case.

The major conclusion we draw from these performance studies is that nearly all the time spent in our new protocol is in the layers concerned with physically transporting messages to a remote machine. Our protocol imposes little cost in relation to this number. This result has convinced us that our new system should be built over some sort of extensible kernel, so that the protocol can be moved closer to the hardware communication device. For example, both Mach and Chorus permit application developers to move modules of code into the network communication component of the kernel. In our case, this would yield a significant speedup. The other obvious speedup would result from the use of hardware multicast, an idea that we are now exploring experimentally.
9 Conclusions

We have presented a protocol efficiently implementing a reliable, causally ordered multicast primitive. The protocol is easily extended into a totally ordered “atomic” multicast primitive, and has been implemented as part of Version 2.1 of the Isis Toolkit. Our protocol offers an inexpensive way to achieve the benefits of virtual synchrony. It is fast and scales well; in fact, there is no evident limit to the size of network in which it could be used. Even in applications with large numbers of overlapping groups, the overhead on a multicast is typically small, and in systems with bursty communication, most multicasts can be sent with no overhead other than that needed to implement reliable, FIFO interprocess channels. With appropriate device drivers or multicast communication hardware, the basic protocol will operate safely in a completely asynchronous, streaming fashion, never blocking a message or its sender unless out-of-order reception genuinely occurs. Our conclusion is that systems such as Isis can achieve performance competitive with the best existing multicast facilities - a finding contradicting the widespread concern that fault-tolerance may be unacceptably costly.
Acknowledgements

The authors are grateful to Maureen Robinson, who prepared many of the figures. We also thank Keith Marzullo, who suggested we search for the necessary and sufficient conditions of Section 6.5, and made many other useful comments. Gil Neiger, Tushar Chandra, Robert Cooper, Barry Gleeson, Shivakant Misra, Aleta Ricciardi, and Mark Wood made numerous suggestions concerning the protocols and presentation, which we greatly appreciate.

References


