Message Logging: Pessimistic, Optimistic, Causal and Optimal

Lorenzo Alvisi*
The University of Texas at Austin
Department of Computer Sciences
Austin, TX

Keith Marzullo†
University of California, San Diego
Department of Computer Science and Engineering
La Jolla, CA

Abstract

Message-logging protocols are an integral part of a popular technique for implementing processes that can recover from crash failures. All message-logging protocols require that, when recovery is complete, there be no *orphan processes*, which are surviving processes whose states are inconsistent with the recovered state of a crashed process. We give a precise specification of the consistency property "no orphan processes". From this specification, we describe how different existing classes of message-logging protocols (namely *optimistic*, *pessimistic*, and a class that we call *causal*) implement this property. We then propose a set of metrics to evaluate the performance of message-logging protocols, and characterize the protocols that are *optimal* with respect to these metrics. Finally, starting from a protocol that relies on causal delivery order, we show how to derive optimal causal protocols that tolerate f overlapping failures and recoveries for a parameter $f : 1 \leq f \leq n$.

Index Terms Message logging, optimistic protocols, pessimistic protocols, checkpoint-restart protocols, resilient processes, specification of fault-tolerance techniques.

1 Introduction

Message-logging protocols (for example, [BBG83, PP83, SY85, JZ87, SBY88, SW89, JZ90, VJ94, EZ92]) are popular for building systems that can tolerate process crash failures. These protocols require that each process periodically record its local state and log the messages it received after having recorded that state. When a process crashes, a new process is created in its place: the new process is given the appropriate recorded local state and then it is sent the logged messages in the order they were originally received. Thus, message logging protocols implement an abstraction of

*This material is based on work supported in part by the Office of Naval Research under contract N00014-91-J-1219, the National Science Foundation under Grant No. CCR-9003440, DARPA/NSF Grant No. CCR-9014363, NASA/DARPA grant NAG-2-893, and AFOSR grant F49620-94-1-0198. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not reflect the views of these agencies.

†This author was supported by the Defense Advanced Research Projects Agency (DoD) under NASA Ames grant number NAG 2–593, Contract N00140-87-C-8904 and by AFOSR grant number F49620-93-1-0242. 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.

a resilient process in which the crash of a process is translated into intermittent unavailability of that process.

All message-logging protocols require that once a crashed process recovers, its state is consistent with the states of the other processes. This consistency requirement is usually expressed in terms of *orphan processes*, which are surviving processes whose states are inconsistent with the recovered state of a crashed process. Thus, message-logging protocols guarantee that upon recovery, no process is an orphan. This requirement can be enforced either by avoiding the creation of orphans during an execution, as *pessimistic* protocols do, or by taking appropriate actions during recovery to eliminate all orphans as *optimistic* protocols do.

This paper examines more carefully the different approaches to message-logging protocols. To do so, we characterize more precisely the property of having no orphans. This characterization is surprisingly simple yet it is general enough to describe both pessimistic and optimistic message-logging protocols. Both classes of protocols, however, have drawbacks, which are apparent from their characterizations. Thus, we use the same characterization to define a third class of protocols that we call *causal* message-logging protocols. These protocols are interesting because they do not suffer from the drawbacks of the other two classes and can be made optimal with respect to a set of reasonable performance metrics.

Finally, starting from a simple but inefficient message-logging protocol that uses causal multicast, we derive an optimal causal protocol that maintains enough information to tolerate f overlapping failures, where f is a parameter of the protocol. We only derive this protocol far enough to illustrate how piggybacking can be used to implement causal message-logging. A thorough discussion of such optimal causal protocols and of the tradeoffs involved in their implementation is presented in [AM96].

The characterization that we give is concerned with how information can be lost in the system due to crashes, and how message-logging protocols cope with this loss of information. There are other issues of message-logging protocols that are out of the scope of this paper, such as communication with the environment, checkpointing, and recovery. A discussion of these issues in the context of causal message logging is presented in [Alv96].

The paper proceeds as follows. Section 2 describes the system model commonly assumed for message-logging protocols. Section 3 discusses the notion of consistency in message-logging protocols. Section 4 presents the derivation of the *always-no-orphans condition*, and explains how it relates with the consistency conditions implemented by pessimistic, optimistic and causal protocols. Sections 5 defines optimal message-logging protocols. In Section 6 we present a simple non-optimal protocol that uses causal delivery to implement the always-no-orphans consistency condition. In Section 6.3 we refine this protocol to obtain an optimal causal message-logging protocol. Section 7 concludes the paper.

2 System Model

Consider a distributed system $\mathcal{N}$ consisting of n processes. Processes communicate only by exchanging messages. The system is *asynchronous*: there exists no bound on the relative speeds of processes, no bound on message transmission delays, and no global time source.

Processes execute events, including *send* events, *receive* events, and local events. These events are ordered by the irreflexive partial order *happens before* $\rightarrow$ that represents potential causality [Lam78]. We assume that a send event specifies only a single destination, but the following could be easily extended to include multiple destination messages. We also assume that the channels between processes do not reorder messages and can fail only by transiently dropping messages.

Processes can fail independently according to the fail-stop failure model [Sch84]. In this model, processes fail by halting, and the fact that a process has halted is eventually detectable by all surviving processes.

Outside of satisfying the happens-before relation, the exact order a process executes receive events depends on many factors, including process scheduling, routing, and flow control. Thus, a recovering process may not produce the same run upon recovery even if the same set of messages are sent to it since they may not be redelivered in the same sequence as before. We represent this low-level nondeterminism by creating a nondeterministic local event *deliver* that corresponds to the delivery of a received message to the application. When an deliver event is executed, any received message can be chosen to be delivered to the application. The only constraints are that a message can be delivered only once, channel order is preserved, and as long as a process does not fail, it eventually delivers all messages that it receives.

A deliver event assigns a *receive sequence number* to a message m that encodes the order in which m was delivered[1]. We denote the receive sequence number of a message m as $m.rsn$. Thus, if process p delivers m and $m.rsn = \ell$ then m is the ℓ^{th} message that p delivered [SY85].

An execution of the system is represented by a *run*, which is the sequence of states the system passes through during that execution. Each state consists of the individual process states, each of which is a mapping from program variables and implicit variables (such as program counters) to values. The state of a process does not include the variables defined in the underlying communication system, such as the queues of messages that have been received but not yet delivered to processes. We assume that only one process changes state between any two adjacent states in the run. Thus, each pair of adjacent states defines an event that was executed by a process. The resulting sequence of events is consistent with the happens before relation.

A *property* is a logical expression evaluated over runs. We use the two temporal operators $\Box$ and $\Diamond$ [Pnu77] to express properties. The property $\Box\Phi$ (read "always Φ") at a state in a run means that the property Φ holds in the current state and all subsequent states of the run. The property $\Diamond\Phi$ (read "eventually Φ") at a state in a run means that the property Φ holds in the current state or in a subsequent state of the run. If no state and run are specified, then a property holds if it holds in the initial state of all runs of the system. For example, the property $\Box((a = 1) \Rightarrow \Diamond(b = 1))$ means that in all runs, for every state in a run, if the state satisfies $a = 1$, then the state or a subsequent state in the run satisfies $b = 1$.

3 Consistency in Message-Logging Protocols

In message-logging protocols, each process typically records both the content and receive sequence number of all the messages it has delivered into a location (called a *message log*) that will survive the failure of the process [Gra77]. This action is called *logging*. To trim message logs, a process may also periodically create a *checkpoint* of its local state. For example, in some message-logging protocols once a process p checkpoints its state, all messages delivered before this state can be removed from p's message log. Note that the periodic checkpointing is only needed to bound the length of message logs (and hence the recovery time of processes). For simplicity we ignore checkpointing in this paper.

Logging a message may take time: therefore, there is a natural design decision of whether or not a process should wait for the logging to complete before sending another message. For example,

[1] A more logical name would be the *deliver sequence number*. We use *receive sequence number* in order to be consistent with message logging literature.

suppose that having delivered message m, process p sends message m' to process q, and q delivers m'. If message m is not logged by the time p sends m', then the crash of p may cause information about m to be lost. When a new process p is initialized and replayed logged messages, p may follow a different run in which it does not send m' to q. In this case, process q would no longer be consistent with p, because q would have delivered a message that was not sent by the current process p. Such a process q is called an *orphan*. Protocols that can create orphans are called *optimistic* because they are willing to take the (hopefully) small risk of creating orphans in exchange of better performance during failure-free runs. If a failure occurs, though, an optimistic protocol determines whether there are any orphans. If so, it rolls them back in order to make the states of the processes consistent again.

A *pessimistic* protocol is one in which no process p ever sends a message m' until it knows that all messages delivered before sending m' are logged. Pessimistic protocols never create orphans, and so reconstructing the state of a crashed process is straightforward as compared to optimistic protocols. On the other hand, pessimistic protocols potentially block a process for each message it receives. This can slow down the throughput of the processes even when no process ever crashes.

There are message-logging protocols that do not exhibit the above tradeoff: they neither create orphans when there are failures nor do they ever block a process when there are no failures. We call such protocols *causal* message-logging protocols, for reasons that will become clear later. Examples of this class are presented in [EZ92] and [AHM93].

All message-logging protocols must address the issue of processes that communicate with the environment. It is natural to model such communications in terms of the sending and delivery of messages. Because one end of the communications is not a process, recovery of a process with respect to these kinds of messages is usually done differently than with respect to messages between processes. For input, the data from the environment must be stored in a location that is accessible for the purpose of replay. For output, a process must be in a recoverable state before sending any message to the environment. This means that, in general, even optimistic message-logging protocols may block before sending a message to the environment. Such issues are outside of the scope of this paper.

4 Specification of Message Logging

Let ρ be a run of $\mathcal{N}$. If no process executes nondeterministic events in ρ and a process fails, then recovering the system to a consistent state is straightforward: the system can stop and restart from its initial state. Since ρ contained only deterministic events, ρ would be re-executed. If processes can execute nondeterministic events in ρ, then some mechanism is needed to guarantee that the same events executed in ρ are again executed during recovery. Message logging protocol implement such a mechanism by recording the information necessary to deterministically re-execute events of ρ.

For each message m delivered during ρ, let $m.source$ and $m.ssn$ denote, respectively, the identity of the sender process and a unique identifier assigned to m by the sender. The latter may, for example, be a sequence number. Let $deliver_{m.dest}(m)$ denote the event that corresponds to the delivery of message m by process $m.dest$, and let $m.data$ be the application data carried by the message. The tuple $\langle m.source, m.ssn, m.dest, m.rsn, m.data\rangle$ determines m and the order in which m was delivered by $m.dest$ relative to the other messages delivered by $m.dest$. We call this tuple the *determinant* of event $deliver_{m.dest}(m)$, and we denote the tuple as $\#m$. For simplicity, we also refer to $\#m$ as the determinant of m.

Therefore, the ability of any message logging protocol to correctly recover from process failures experienced during ρ depends crucially on the availability upon recovery of the determinants of the delivery events that occurred during ρ.

However, the immediate availability of the content $m.data$ of a message m exchanged during ρ is not an essential part of $\#m$. In practice, an effective technique for logging the content of messages can result in dramatic performance gains, but, at least in principle, the content of a message can be deterministically regenerated from the initial states of the processes and the determinants of each deliver event in ρ. In other words, logging messages is not an essential task of message logging protocols, while logging message determinants is. Hence, we will ignore in our specification the specifics of logging the contents of messages, just as we ignore the details related to taking periodic checkpoints— both are caching that can be added to speed up recovery.[2]

Our goal is to derive a condition, which we call *always-no-orphans*, that guarantees that no orphans are generated during any run of $\mathcal{N}$. To do so, we define two subsets of $\mathcal{N}$ for each message m delivered in a run. The first set, $Depend(m)$, contains all processes whose state reflects delivery of m: $Depend(m)$ contains the destination of message m, plus any process that delivered a message sent causally after the delivery of m. Formally,

$$Depend(m) \stackrel{\text{def}}{=} \left\{ j \in \mathcal{N} \;\middle|\; \begin{array}{ll} \vee & ((j = m.dest) \wedge j \text{ has delivered } m) \\ \vee & (\exists m' \colon (deliver_{m.dest}(m) \rightarrow deliver_j(m'))) \end{array} \right\}$$

The second set, $Log(m)$, is used to characterize the set of processes that have a copy of $\#m$ in their volatile memory.[3] Process $m.dest$ becomes a member of $Log(m)$ when it delivers m.

Suppose that a set of processes $\mathcal{C} \subseteq \mathcal{N}$ fail. The determinant of a deliver event $deliver_{m.dest}(m)$ is lost if $Log(m) \subseteq \mathcal{C}$. Assume that $\#m$ is lost. By definition, process $m.dest$ was in $Log(m)$ before the failure. Therefore, $m.dest$ must be in $\mathcal{C}$ and so must be recovered. Since $\#m$ is lost, $m.dest$ may not be able to deliver m in the correct order during recovery. Consequently, $m.dest$ may not be able to regenerate some of the messages it sent after executing $deliver_{m.dest}(m)$ in the original execution. All processes whose state depends on process $m.dest$ delivering message m in the correct order will become orphan processes.

We say that a process p becomes an orphan of $\mathcal{C}$ when p itself does not fail and p's state depends on the delivery of a message m whose determinant has been lost. Formally:

$$p \text{ } orphan \text{ } of \text{ } \mathcal{C} \stackrel{\text{def}}{=} \left(\begin{array}{ll} \wedge & (p \in \mathcal{N} - \mathcal{C}) \\ \wedge & \exists m : ((p \in Depend(m)) \wedge (Log(m) \subseteq \mathcal{C})) \end{array} \right) \tag{1}$$

Negating (1) and quantifying over p gives the following necessary and sufficient condition for there being no orphans created by the failure of a set of processes $\mathcal{C}$:

$$\forall m : ((Log(m) \subseteq \mathcal{C}) \Rightarrow (Depend(m) \subseteq \mathcal{C})) \tag{2}$$

Our goal is to derive a property that guarantees that no set $\mathcal{C}$ of faulty processes results in the creation of orphans. Quantifying (2) over all $\mathcal{C}$, we obtain:

$$\forall m : (Depend(m) \subseteq Log(m)) \tag{3}$$

[2]A similar argument can be made about $m.dest$ not being essential in the determinant of m. We keep $m.dest$ in $\#m$ for ease of exposition.

[3]One can imagine protocols where no single process knows the value of $\#m$ but a set of processes collectively do. An example of such a protocol is given in [JV87], where the value of $\#m$ for some message m may be inferred by "holes" in the sequence of logged receive sequence numbers. Since this is the only such case that we know of and it can tolerate at most two process failures at a time, we do not consider further this more general method of logging.

Since we want (3) to hold for every state, we require the following property

$$\forall m : \Box(\mathit{Depend}(m) \subseteq \mathit{Log}(m)) \tag{4}$$

Property (4) is a safety property: it must hold in every state of every execution to ensure that no orphans are created. It states that, to avoid orphans, it is sufficient to guarantee that if the state of a process p depends on the delivery of message m, then p must keep a copy of m's determinant in its volatile memory.

We say that $\#m$ is *stable* (denoted $\mathit{stable}(m)$) when $\#m$ cannot be lost. Condition 3 must hold only for messages with a determinant that is not stable. Hence 3 need only hold when $\#m$ is not stable:

$$\forall m : \Box(\neg \mathit{stable}(m) \Rightarrow (\mathit{Depend}(m) \subseteq \mathit{Log}(m))) \tag{5}$$

We call this property the *always no orphans condition.* If determinants are kept in stable storage [Gra77], then $\mathit{stable}(m)$ holds when the write of $\#m$ to stable memory completes. If determinants are kept in volatile memory, and we assume that no more than f processes — where $|\mathcal{C}| \leq f$ — can fail concurrently, then $\mathit{stable}(m)$ holds as long as $f+1$ processes have a copy of $\#m$ in their volatile memory. In the latter case, the *always no orphans condition* can be written:

$$\forall m : \Box((|\mathit{Log}(m)| \leq f) \Rightarrow (\mathit{Depend}(m) \subseteq \mathit{Log}(m))) \tag{6}$$

4.1 Pessimistic and Optimistic Protocols Revisited

As we saw in Section 3, pessimistic message-logging protocols do not create orphans. Therefore, they must implement Property (5). Indeed, they implement the following stronger property:

$$\forall m : \Box(\neg \mathit{stable}(m) \ \Rightarrow (|\mathit{Depend}(m)| \leq 1)) \tag{7}$$

In practice, (7) does not allow the process that delivers message m to send any messages until the determinant of m is stable.[4] To see this, suppose process p_1 has just delivered m. $\mathit{Depend}(m)$ contains only p_1, so $|\mathit{Depend}(m)| = 1$. Suppose now that p_1 sends a message m' to process p_2. Since $\mathit{deliver}_{p_1}(m) \rightarrow \mathit{deliver}_{p_2}(m')$, when process p_2 delivers m' it becomes a member of $\mathit{Depend}(m)$, and $|\mathit{Depend}(m)| = 2$. Thus, as soon as m' is sent, the consequent in (7) can become false. To preserve (7), p_1 ensures that $\mathit{stable}(m)$ holds before sending m'. If the time during which $\neg\mathit{stable}(m)$ is brief, then it is unlikely that p will attempt to send a message while $\#m$ is not stable, so any performance loss by inhibiting message sends is (hopefully) small.

The reasoning behind optimistic protocols starts from the same assumption used by pessimistic protocols: the time during which $\neg\mathit{stable}(m)$ holds is brief. Hence, (5) holds trivially nearly all the time. Therefore, it is not practical to incur the performance cost that pessimistic protocols suffer. Instead, optimistic protocols take appropriate actions during recovery to re-establish (5) in the unlikely event that it is violated as a result of the failure of a set of processes $\mathcal{C}$.

Optimistic protocols implement the following property:

$$\forall m : \Box(\neg \mathit{stable}(m) \ \Rightarrow ((\mathit{Log}(m) \subseteq \mathcal{C}) \Rightarrow \Diamond(\mathit{Depend}(m) \subseteq \mathcal{C}))) \tag{8}$$

Property (8) is weaker than Property (2), and therefore weaker than Property (5). Property (8) permits the temporary creation of orphan processes, but guarantees that, by the time recovery

[4] In pessimistic sender-based logging [JZ87] process $m.\mathit{dest}$ increases $|\mathit{Log}(m)|$ by sending the value of $m.\mathit{rsn}$ to process $m.\mathit{source}$, piggybacked on the acknowledgment of message m. Notice however that $m.\mathit{dest}$ is still not allowed to send any *application* messages until it is certain that $m.\mathit{source}$ has become a member of $\mathit{Log}(m)$.

is complete, no surviving process will be an orphan and Property (2) will hold. This is achieved during recovery by rolling back orphan processes until their states do not depend on any message whose determinant has been lost. In other words, $Depend(m)$ is made smaller (or, equivalently, $\mathcal{C}$ is made larger) until $Depend(m) \subseteq \mathcal{C}$ and Property (2) is restored.

4.2 Causal Message-Logging Protocols

Pessimistic protocols implement a property stronger than Property (4) and therefore never create orphans. However, implementing this stronger property may introduce blocking in failure-free runs. Optimistic protocols avoid this blocking by satisfying a property weaker than Property (4). It is natural to ask whether a protocol can be designed that implements Property (4) — and therefore creates no orphans—yet does not implement a property as strong as Property (7)—and thus does not introduce any blocking. We find such a property by strengthening Property (6).

Consider the following property:

$$\forall m : \Box((|Log(m)| \leq f) \Rightarrow (Depend(m) = Log(m))) \tag{9}$$

Property (9) strengthens (6), and so protocols that implements (9) prevents orphans. Furthermore, such protocols disseminate the least number of copies of $\#m$ needed in order to satisfy (6), thereby conserving storage and network bandwidth.

Unfortunately, satisfying Property (9) requires processes to be added to $Log(m)$ and $Depend(m)$ simultaneously. Satisfying this requirement would result in complicated protocols. Thus, we consider a different strengthening of Property (6) that is weaker than Property (9) but still bounds $Log(m)$:

$$\forall m : \Box((|Log(m)| \leq f) \Rightarrow ((Depend(m) \subseteq Log(m)) \ \wedge \Diamond(Depend(m) = Log(m)))) \tag{10}$$

This characterization strongly couples logging with causal dependency on deliver events. It requires that:

- As long as $\#m$ is at risk of being lost, all processes that have delivered a message sent causally after the delivery of m have also stored a copy of $\#m$.
- Eventually, all the processes that have stored a copy of $\#m$ deliver a message sent causally after the delivery of m.

We call the protocols that implement Property (10) *causal* message-logging protocols. Family-Based-Logging [AHM93] and Manetho [EZ92] are two examples of causal message-logging protocols for the special cases $f = 1$ and $f = n$ respectively.

5 Optimal Message-Logging Protocols

We now consider four metrics with which one can compare different message-logging protocols. These metrics are not the only ones that one might wish to consider when making such a comparison, but they do give some measure of the running time of the protocol, both in the failure-free case and during recovery.

To express these four metrics, let Π be a protocol executed by a set $\mathcal{N}$ of processes, and assume that Π is written to tolerate no process failures and no transient channel failures. Let Π_μ denote Π combined with a message-logging protocol μ that tolerates process failures and transient channel failures.

1. *Number of forced roll-backs*: This metric gives some measure of the amount of work that might need to be discarded as a result of a process crash. Consider a run ρ of Π_μ, and suppose that a set $\mathcal{C} \subseteq \mathcal{N}$ of processes fails. We say that μ forces r roll-backs if there is a run in which μ requires r correct processes to roll back their state, and for all ρ, μ requires no more than r correct processes to roll back their state. For example, optimistic sender-based logging [JZ87] forces $|\mathcal{N}|$ rollbacks.

   A lower bound for r is 0. The existence of pessimistic protocols, where no correct process ever rolls back, establishes that this bound is tight.

2. *k–blocking*: This metric gives some measure of the amount of idle time that can be added to the execution of a process during a failure-free run. Consider a message m delivered by process $m.dest$, and let e be the first send event of process $m.dest$ that causally follows $deliver_{m.dest}(m)$. We say that a message-logging protocol is *k–blocking* if, in all failure-free runs and for all messages m, process $m.dest$ delivers no less than k messages between $deliver_{m.dest}(m)$ and e. For example, pessimistic sender-based logging [JZ87] is 1–blocking because process $m.dest$ must receive a message acknowledging the logging of $m.rsn$ before sending a message subsequent to the delivery of m. Optimistic protocols are, by design, 0–blocking. For completeness, we define pessimistic protocols that log the determinant of m in local stable storage to be 1–blocking, since there is one event (the acknowledgment of $\#m$ being written to stable storage) that must occur between $deliver_{m.dest}(m)$ and a subsequent send [BBG83].

   A lower bound for k is 0. The existence of optimistic protocols establishes that this bound is tight.

3. *Number of messages*: This metric gives one measure of the load on the network generated by the message logging protocol. Protocol Π can be transformed into a protocol Π_ℓ that tolerates transient channel failures using an acknowledgment scheme: whenever process $m.dest$ receives a sequence of messages generated by Π, it sends an acknowledgment to process $m.source$. Process $m.source$ resends a message until it receives such an acknowledgment. In all failure-free runs, including failure-free ones, protocol Π_ℓ will send more messages than Π.

   Suppose now that ρ is run using Π_μ instead of Π_ℓ. We say that μ sends additional messages in ρ if Π_μ sends more messages than Π_ℓ. For example, in order to tolerate single crash failures pessimistic sender-based logging [JZ87] potentially requires that one extra acknowledgment be sent for each application message sent.

   A lower bound is to send no additional messages. The existence of optimistic protocols establishes that this bound is tight.

4. *Size of messages*: This metric gives another measure of the load on the network generated by the message logging protocol. Consider any message m sent in a run of Π_ℓ, and let m_μ be the corresponding message sent in the equivalent failure-free run of Π_μ. Let $|m|$ and $|m_\mu|$ be the size of m and m_μ, respectively. We will say that μ sends a additional data if the maximum value of $|m_\mu| - |m|$ is a for all such m and m_μ taken from all pairs of corresponding runs. For example, for optimistic protocols that track direct dependencies, a is a constant [SW89, VJ94]; while for optimistic protocols that track transitive dependencies a is proportional to the number of processes in the system [SY85, SW89].

   A lower bound for a is 0. The existence of pessimistic receiver-based logging proves that this bound is tight.

There is a tradeoff between the number of additional messages sent in Π_μ and the sizes of these messages: a message-logging protocol μ could include less information in a message by sending additional messages containing the extra data. In the majority of deployed computer networks there is, to a point, a performance benefit in sending a few large messages instead of several small messages. Hence, we prefer to keep the number of additional messages small at the expense of message size, and we define an *optimal message-logging protocol* as a message-logging protocol that is 0–blocking, introduces 0 forced roll-backs, and sends no additional messages. It is a trivial exercise to change protocols that are optimal in this sense to protocols that minimize the size of messages by sending additional messages.

Of course, protocols that are optimal according to our definition do not necessarily outperform in practice non-optimal protocols. Other issues that are difficult to quantify, such as the cost of output commit, must be taken into consideration in assessing the performance of a message-logging protocol [EZ94]. Nevertheless, the protocols we call optimal are unique in optimally addressing the theoretical *desiderata* of the message-logging approach.

6 Using Causal Delivery Order to Enforce the Always-No-Orphans Consistency Condition

We now derive an optimal causal message-logging protocol, i.e. an optimal message-logging protocol that implements Property (10). We do so by first presenting, in this section, a simple non-optimal protocol that uses causal delivery order to implement the always-no-orphans consistency condition. In the following section, we refine this protocol and obtain an optimal causal message-logging protocol. We begin by defining causal delivery order.

6.1 Causal Delivery Order

Let $send_p(m)$ to q denote the event whereby process p sends message m to process q, and $receive_q(m)$ the event whereby q receives m.

FIFO delivery order guarantees that if a process sends a message m followed by a message m' to the same destination process, then the destination process does not deliver m' unless it has previously delivered m. Formally:

$$\begin{aligned} &\forall p, r, m, m' : send_p(m) \text{ to } r \rightarrow send_p(m') \text{ to } r \\ &\quad \Rightarrow deliver_r(m) \rightarrow deliver_r(m') \end{aligned} \tag{11}$$

FIFO delivery order constrains the order in which the destination process delivers messages m and m' only when m causally precedes m' and m and m' are sent by the same process. Causal delivery order [BJ87] strengthens FIFO delivery order by removing the requirement that ordering occurs only when the source of m and m' are the same. It guarantees that if the sending of m causally precedes the sending of m' and m and m' are directed to the same destination process, then the destination process does not deliver m' unless it has previously delivered m. Formally:

$$\begin{aligned} &\forall p, q, r, m, m' : send_p(m) \text{ to } r \rightarrow send_q(m') \text{ to } r \Rightarrow \\ &\quad deliver_r(m) \rightarrow deliver_r(m') \end{aligned} \tag{12}$$

An example of an execution where process r delivers messages according to causal delivery order is shown in Figure 1. Since $send_p(m_3)$ to r causally follows $send_q(m_1)$ to r, r delivers m_3 only after m_1 has been delivered.

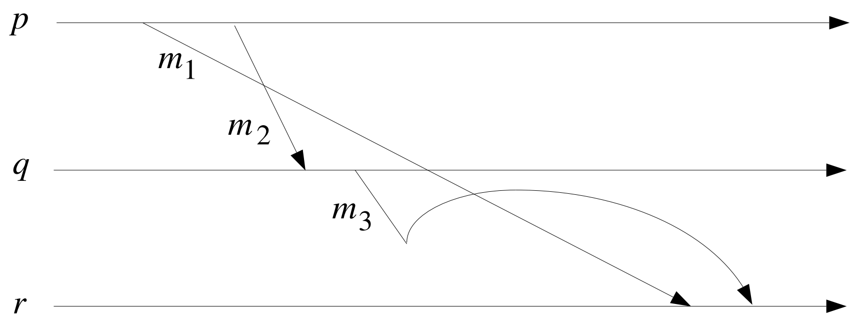

Figure 1: An example of causal delivery order

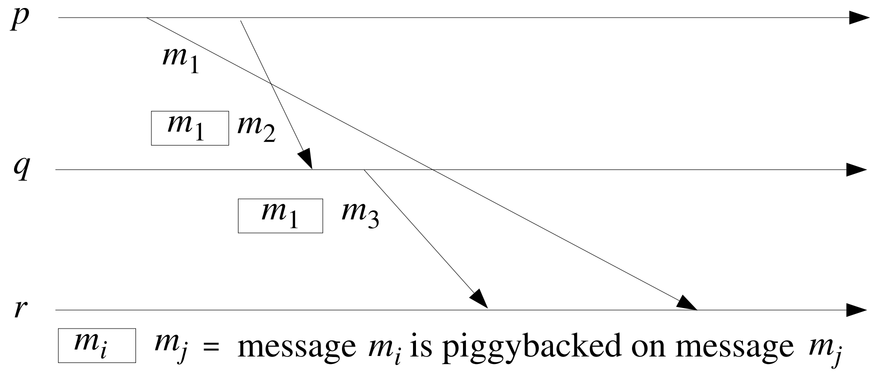

Figure 2: Implementing causal delivery order through piggybacking

There are two fundamental approaches to implementing causal message delivery. The first is to add to each message m additional information that m's destination process uses to determine when m can be delivered. [RST91, SS92, BSS91] Using this approach, process r in Figure 1 would realize, when it receives m_3, that it must wait to receive m_1 and would delay delivery of m_3 accordingly. The problem with this approach is that slow messages can significantly affect the performance of the system, since they prevent faster messages from being delivered.

The second approach is for each process p to piggyback, on each message m that p sends, all messages m' sent in the causal history of the event $send_p(m)$ such that p does not know if m' has already been delivered [BJ87]. The piggybacked messages are placed in a total order that extends the partial order imposed by the happens-before relation. Before delivering m, process $m.dest$ first checks if any message m' in m's piggyback has $m.dest$ for destination. If so, p delivers m only after each such message m' has been delivered according to causal delivery order. This is the approach illustrated in Figure 2. Message m_1 is piggybacked on m_2 and m_3, since p and q do not know whether m_1 has already been delivered. When r receives m_3, it checks the piggyback to find m_1. Process r therefore immediately delivers m_1 and then delivers m_3 without waiting for the slower copy of message m_1 to arrive. When the copy of m_1 is received directly from q, r discards it.

The problem with this piggybacking approach is that, even though several optimizations can be used to reduce the size of the piggyback, the overhead on message size can become very large.

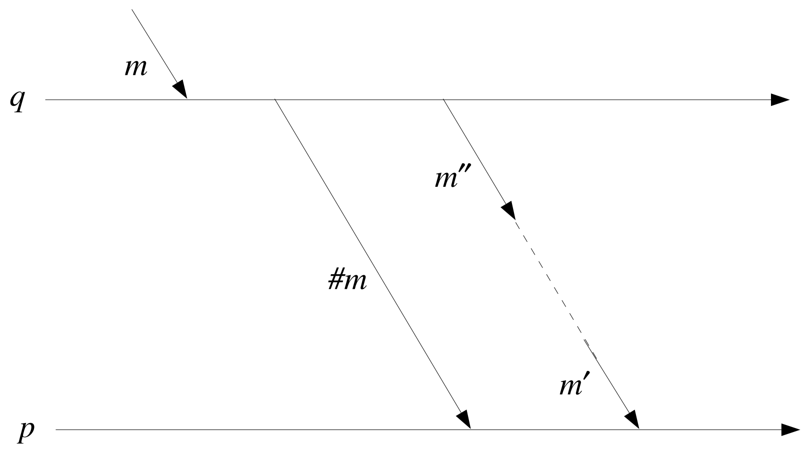

Figure 3: $Depend(m) \subseteq Log(m)$ in Π_{cd}

6.2 A Suboptimal Causal Protocol

We now present a Protocol Π_{cd} that uses causal delivery order to satisfy Property (4). Processes in Π_{cd} behave as follows:

(i) Processes exchange two kinds of messages: *application messages* and *determinant messages*. A determinant message contains the determinant of an application message.

(ii) Application and determinant messages are delivered according to causal delivery order.

(iii) Suppose process p receives an application message m. To deliver m, p creates the determinant $\#m$ of message m and logs $\#m$ in its volatile memory. Then, p supplies $m.data$ to the application.

(iv) Suppose process p delivers an application message m. Before sending any subsequent application messages, p sends a determinant message containing $\#m$ to all the other processes.

(v) Suppose process p receives a determinant message containing $\#m$. To deliver the determinant message, p logs $\#m$ in its local volatile memory.

Theorem 1 *Protocol Π_{cd} satisfies Property (4).*

Proof. Property (4) requires that

$$\forall m : ((Depend(m) \subseteq Log(m))$$

hold throughout execution. According to the definition of $Depend(m)$ given in Section 4, a process p is a member of $Depend(m)$ for an application message m if one of the following two cases holds:

Case 1: p is the destination of m.

Case 2: p is the destination of an application message m', and $deliver_{m.dest}(m)$ is in the causal history of $deliver_p(m')$.

We now show that in both cases protocol Π_{cd} guarantees that if $p \in Depend(m)$ then $p \in Log(m)$, so $Depend(m) \subseteq Log(m)$ holds.

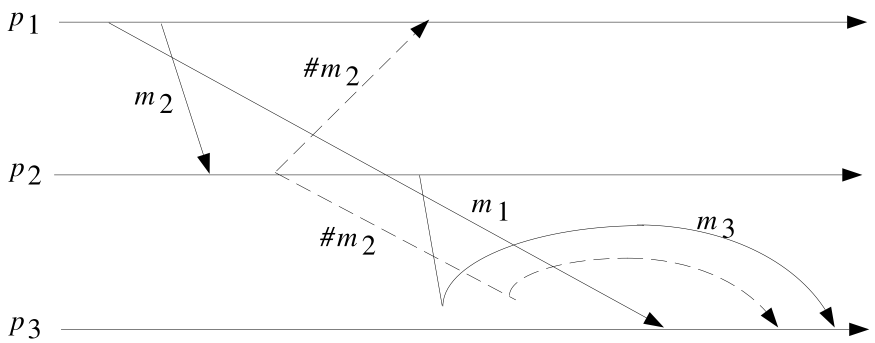

Figure 4: An execution of Π_{cd}

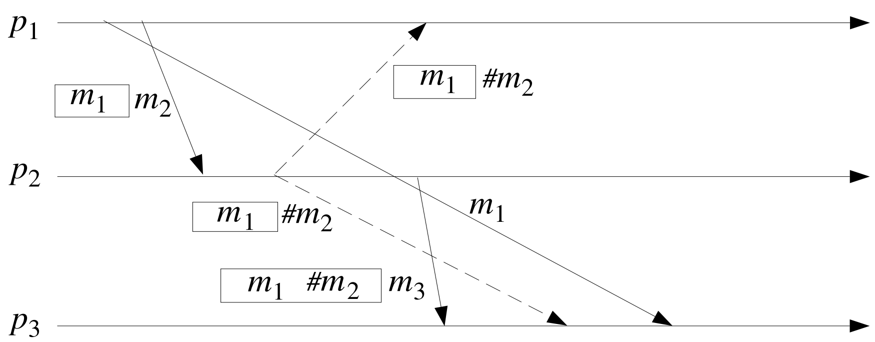

Figure 5: Implementing causal delivery order through piggybacking

Case 1: By point (iii) of Π_{cd}, if p is the destination of m, then p will log $\#m$ in its volatile storage before delivering m. Hence, $Depend(m) \subseteq Log(m)$ holds.

Case 2: If p is not the destination of m, then some other process $q \neq p$ delivered m, and p delivered a message m' such that $deliver_q(m) \rightarrow deliver_p(m')$. Furthermore, there must exist an application message m'', not necessarily distinct from m', such that $deliver_q(m) \rightarrow send_q(m'') \rightarrow deliver_p(m')$ (see Figure 3). By point (iv) of Π_{cd}, q must have sent a determinant message containing $\#m$ to all processes — including process p — before sending m''. By Point (ii) of Π_{cd}, all messages are delivered according to causal delivery: therefore p must have delivered the message containing $\#m$ before delivering m''. It follows from Point (v) of Π_{cd} that p must have logged $\#m$ in its volatile storage before delivering m''. Hence, $p \in Log(m)$ holds. □

Figure 4 shows an execution of Protocol Π_{cd}. Process p_1 sends application message m_1 to process p_3, and then sends application message m_2 to process p_2. After delivering m_2, and before sending application message m_3, p_2 sends to all processes a determinant message containing $\#m_2$. Note that process p_3 first receives m_3, then $\#m_2$, and finally m_1. However, in order to respect causal delivery order, p_3 actually delivers first m_1, then $\#m_2$, and finally m_3.

6.3 An Optimal Causal protocol

Protocol Π_{cd} has several limitations. First, it is not optimal, since for each deliver event of an application message m, it uses additional determinant messages to send $\#m$ to all processes. Sec-

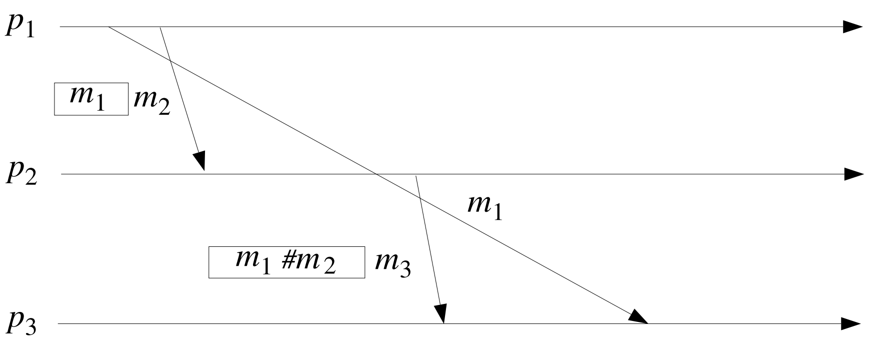

Figure 6: Determinants are piggybacked on application messages

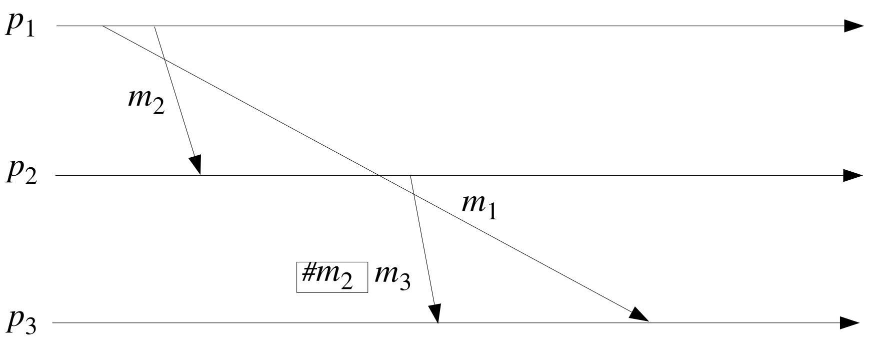

Figure 7: An execution of Π_{oc}

ond, it does not satisfy Property (10), since a process can enter $Log(m)$ by receiving a copy of $\#m$ without ever becoming a member of $Depend(m)$. Finally, it forces both application and determinant messages to be delivered according to causal delivery order, even though causal delivery order is necessary only to regulate how determinant messages are delivered with respect to application messages.

However, by choosing an appropriate implementation of causal delivery order, protocol Π_{cd} can serve as the starting point for a more efficient protocol. Since we desire a non-blocking protocol, our first step is to choose an implementation of causal delivery based on the piggybacking scheme described in Section 6.1. Figure 5 shows the effects of using the piggybacking scheme on the execution in Figure 4.

Observe that, since for each application message m the determinant $\#m$ is piggybacked on any application message sent causally after event $deliver_p(m)$, there is no need to explicitly send $\#m$ in a separate determinant message. Instead, we require each process, before delivering a message m, to deliver all the determinants piggybacked on m. The resulting protocol successfully addresses the first two limitations of the original protocol. By piggybacking $\#m$ on application messages, the protocol ensures that no correct processes p will enter $Log(m)$ unless p will eventually join $Depend(m)$. Furthermore, the protocol is optimal in the number of messages it sends because the determinants are piggybacked on existing application messages. Figure 6 shows the effect of the new protocol on the execution in Figure 5.

Note that, once the piggybacking of the determinants is in place, we are guaranteed that no process will deliver an application message m unless it first delivers the determinants of all the messages delivered in the causal history of $send_{m.source}(m)$. This is precisely the property that we need to guarantee that $\forall m : \mathit{Depend}(m) \subseteq \mathit{Log}(m)$ holds. Therefore, provided that we piggyback the determinants, we can safely relax the requirement that application messages be delivered using causal delivery order. Figure 7 shows the effects of the new protocol on the execution in Figure 6.

The last step of the derivation is to adapt the above protocol to the case where we assume that no more than f processes fail concurrently. In this case, a process piggybacks $\#m$ on an application message only as long as $\#m$ is not stable — i.e. as long as $|\mathit{Log}(m)| \leq f$. We call the resulting protocol Π_{oc}. Processes in Π_{oc} behave as follows:

(i) Processes exchange only application messages. However, determinants may be piggybacked on application messages.

(ii) Suppose process p sends a message m to process q. p piggybacks on m all the non-stable determinants $\#m'$ such that p has $\#m'$ its volatile memory.

(iii) Suppose process p receives a message m. Before delivering m, process p first logs all the determinants piggybacked on m in its determinant log, which is implemented in p's volatile memory. Then, p creates the determinant for message m, and logs it in the determinant log. Finally, p delivers m by supplying $m.data$ to the application.

Theorem 2 *Protocol Π_{oc} satisfies Property (10). Furthermore, Protocol Π_{oc} is optimal.*

Proof. We first prove that Π_{oc} satisfies Property (10). In particular, we prove that:

(a) $(|\mathit{Log}(m)| \leq f) \Rightarrow (\mathit{Depend}(m) \subseteq \mathit{Log}(m))$

(b) $(|\mathit{Log}(m)| \leq f) \Rightarrow \Diamond(\mathit{Depend}(m) = \mathit{Log}(m))$

are satisfied by executions of Π_{oc}. We first observe that, if $|\mathit{Log}(m)| > f$ then (a) and (b) are trivially true. Therefore we consider the case in which $|\mathit{Log}(m)| \leq f$.

We prove (a) by showing that, whenever $|\mathit{Log}(m)| \leq f$ holds, if a process p is a member of $\mathit{Depend}(m)$, then p is also a member of $\mathit{Log}(m)$. Process p is a member of $\mathit{Depend}(m)$ if either p is the destination of m or p is the destination of a message m', and $\mathit{deliver}_{m.dest}(m) \rightarrow \mathit{deliver}_p(m')$.

In the first case, by Point (iii) of Π_{oc}, p must have saved $\#m$ in its determinant log. Therefore, if p is a member of $\mathit{Depend}(m)$, then p must be a member of $\mathit{Log}(m)$.

In the second case, we proceed by induction on the length ℓ of the causal chain of processes associated with a causal path that starts with event $\mathit{deliver}_{m.dest}(m)$ and ends with event $\mathit{deliver}_p(m')$.

Base Case : $\ell = 1$.

If $\ell = 1$, then process $m.dest$ must have been the sender of message m'. In particular, it must be the case that

$$\mathit{deliver}_{m.dest}(m) \rightarrow \mathit{send}_{m.dest}(m') \text{ to } p \rightarrow \mathit{deliver}_p(m').$$

By Point (ii) of Π_{oc}, process $m.dest$ must have piggybacked on m' all the determinants that were not stable at the time of event $\mathit{send}_{m.dest}(m')$ to p. In particular, if $\#m$ was not stable, then it must have been piggybacked on m'. Because of point (iii) of Π_{oc}, if $\#m$ was piggybacked on m', then p must have saved $\#m$ in its determinant log before delivering m'. Therefore, if p is a member of $\mathit{Depend}(m)$, and $|\mathit{Log}(m)| \leq f$, then p is also a member of $\mathit{Log}(m)$.

Inductive Step: We assume that (a) holds for all causal paths whose associated causal chain of processes has length $\ell \leq c$. We prove that (a) holds for any causal path whose associated causal chain of processes has length $\ell = c + 1$.

Consider a causal path that starts with event $deliver_{m.dest}(m)$ and ends with event $deliver_p(m')$, and assume that the associated causal chain of processes has length $\ell = c + 1$. Let q be the process that sent message m' to p. Since, by assumption, a process never sends a message to itself, it follows that $p \neq q$. Therefore, the causal chain associated with the causal path that starts with $deliver_{m.dest}(m)$ and ends with $send_q(m')$ to p must be of length c. By the inductive hypothesis, if $|Log(m)| \leq f$ when m' was sent, then q must have had $\#m$ in its determinant log. By Point (ii) of Π_{oc}, q must have piggybacked $\#m$ on m'. By point (iii) of Π_{oc}, when p received m', p must have added $\#m$ to its determinant log before delivering m. Therefore, if p is a member of $Depend(m)$, and $|Log(m)| \leq f$, then p is also a member of $Log(m)$.

We now prove Part (b), assuming $|Log(m)| \leq f$.

If process p is a member of $Log(m)$, then it must have stored a copy of $\#m$ in its determinant log. For this to happen, p must have either received m, or it must have received a message m', onto which $\#m$ was piggybacked, such that $deliver_{m.dest}(m) \rightarrow send_{m'.source}(m')$. In the first case, unless p fails, it will eventually deliver m and become a member of $Depend(m)$. The second case is similar: unless p fails, it will eventually deliver m' and become a member of $Depend(m)$. Furthermore, since there is only a finite number of processes in $\mathcal{N}$ and $Log(m) \subseteq \mathcal{N}$, eventually all non-faulty members of $Log(m)$ will join $Depend(m)$ and (b) will hold. Note that, if p fails, then p will no longer be a member of $Depend(m)$. However, in this case, p will lose all the data in its determinant log, which is implemented in volatile memory, and will therefore leave $Log(m)$. Hence, (b) holds whether p is a correct process or not. This concludes the proof that Π_{oc} satisfies Property (10).

We now have only to show that Π_{oc} is optimal. To do so, we observe that no additional messages are generated by protocol Π_{oc} over the ones needed by the application, because determinants are piggybacked on existing application messages. Furthermore, no correct processes are forced to rollback as a result of a process' failure, because Π_{oc} is a causal protocol. Finally, Π_{oc} is 0-blocking, because none of (a)-(c) imply blocking. $\Box$

6.4 Implementation Issues

Protocol Π_{oc}'s piggybacking scheme guarantees that all the determinants needed to recover the system to a consistent global state from up to f concurrent failures will be available during recovery. However, the scheme leaves several open questions. First, Π_{oc} does not specify how to collect and use the logged determinants to perform recovery. Furthermore, Π_{oc} assumes that processes have knowledge of the current value of $|Log(m)|$ when they determine whether or not to piggyback the determinant $\#m$ on an application message. This assumption is not realistic in an asynchronous distributed system, since — as we have observed before — it requires processes to have instantaneous access to $Log(m)$, which is defined over the entire distributed system. Since the focus of this paper is not the implementation of optimal causal message-logging protocols but rather their specification, we do not address these issues here. The interested reader is referred to [AM96], in which we discuss the implementation of a class of optimal causal message-logging protocols that tolerate f overlapping failures and recoveries for a parameter $f : 1 \leq f \leq n$. These protocols are based on the same piggybacking scheme given above.

7 Conclusions

In this paper, we have shown that a small amount of simple formalism can go a long way. We have presented the first specification of the message logging approach to fault-tolerance, by precisely defining the *always-no-orphans* condition, which is at the foundation of the consistency requirements of all the protocols in this class. The specification is general enough to describe all existing classes of message logging protocols, yet simple enough to be effectively used as a tool to derive more efficient protocols. We have provided a set of metrics to evaluate the performance of a message logging protocol, and given a set requirements that a protocol must meet in order to be optimal. Finally, we have argued the existence of optimal protocols that tolerate f overlapping failures and recoveries for a parameter $f : 1 \leq f \leq n$ by showing how such a protocol can be derived starting from a simple protocol that relies on causal delivery order. A thorough discussion of optimal protocols and of the tradeoffs involved in their implementation is presented in [AM96, Alv96].

Acknowledgments We would like to thank Bruce Hoppe, Fred Schneider and Sam Toueg for many discussions on this research and on early versions of this paper. We would also like to thank the anonymous referees for their helpful suggestions.

A Derivations

In this section, we show the derivations of the equations in this paper that may not be immediately obvious.

Derivation of Equation (2) from Equation (1)

$$\forall p : \neg(p \ \mathit{orphan\ of}\ \mathcal{C})$$

$$= \langle\!\langle \text{Definition of } p \ \mathit{orphan\ of}\ \mathcal{C} \rangle\!\rangle$$

$$\forall p : \neg \left(\begin{array}{l} (p \in \mathcal{N} - \mathcal{C}) \ \wedge \\ \exists m : ((p \in \mathit{Depend}(m)) \wedge (\mathit{Log}(m) \subseteq \mathcal{C})) \end{array} \right)$$

$$= \langle\!\langle \text{De Morgan's Law} \rangle\!\rangle$$

$$\forall p : \left(\begin{array}{l} (p \notin \mathcal{N} - \mathcal{C}) \ \vee \\ (\forall m : (\neg(p \in \mathit{Depend}(m)) \vee \neg(\mathit{Log}(m) \subseteq \mathcal{C}))) \end{array} \right)$$

$$= \langle\!\langle \text{For } x \text{ not free in Q: } (\forall x : P(x)) \vee Q = \forall x : (P(x) \vee Q) \rangle\!\rangle$$

$$\forall p, m : ((p \in \mathcal{C}) \vee \neg(p \in \mathit{Depend}(m)) \vee \neg(\mathit{Log}(m) \subseteq \mathcal{C}))$$

$$= \langle\!\langle \text{Definition of } \Rightarrow \rangle\!\rangle$$

$$\forall p, m : ((\mathit{Log}(m) \subseteq \mathcal{C}) \Rightarrow ((p \in \mathit{Depend}(m)) \Rightarrow (p \in \mathcal{C})))$$

$$= \langle\!\langle P \subseteq Q = \forall p : (p \in P \Rightarrow p \in Q) \rangle\!\rangle$$

$$\forall m : ((\mathit{Log}(m) \subseteq \mathcal{C}) \Rightarrow (\mathit{Depend}(m) \subseteq \mathcal{C}))$$

Derivation showing that Equation (7) implies Equation (5)

1. $\forall m : \Box(\neg \mathit{stable}(m) \;\Rightarrow (|\mathit{Depend}(m)| \leq 1))$
   $\langle\!\langle$Definition of $\mathit{Depend}(m)\rangle\!\rangle$
2. $\forall m : \Box((|\mathit{Depend}(m)| \leq 1) \Rightarrow (\mathit{Depend}(m) \subseteq \{m.\mathit{dest}\}))$
   $\langle\!\langle$Definition of $\mathit{Log}(m)\rangle\!\rangle$
3. $\forall m : (\{m.\mathit{dest}\} \subseteq \mathit{Log}(m))$
   $\langle\!\langle$Transitivity of Set Inclusion, using 2 and 3$\rangle\!\rangle$
4. $\forall m : \Box(|\mathit{Depend}(m)| \leq 1) \Rightarrow (\mathit{Depend}(m) \subseteq \mathit{Log}(m))$
   $\langle\!\langle$Transitivity of Implication, using 1 and 4$\rangle\!\rangle$
5. $\forall m : \Box(\neg \mathit{stable}(m) \;\Rightarrow (\mathit{Depend}(m) \subseteq \mathit{Log}(m)))$

B Sharing the Log

The specification derived in Section 4 assumes that a process logs determinants either in its own volatile memory or to stable memory. Hence, determinants may be lost only when a process fails, and since processes fail independently, different copies of the same determinant are lost independently. A more general approach is to regard determinant logs as being first-class objects. By doing so, one can model a set of processes sharing storage for logging purposes and thereby decouple the failure of processes from the loss of determinants. For example, one reasonable approach would be to implement the logging component of a message-logging protocol so that a single, shared log is maintained for all the processes that run on the same processor. This would result in a more efficient implementation of causal message logging.

Define a *logging site* to be a storage object that a process can read and write. Each process uses a logging site for its determinant log. We assume that logging sites fail and recover. When a logging site fails, the values written to the logging site are lost, and all processes using that logging site also fail. A process may fail without its associated logging site failing, however.

When determinants are kept at logging sites, $\mathit{Log}(m)$ denotes the set of logging sites that contain the determinant of event $\mathit{deliver}_{m.\mathit{dest}}(m)$. An implementation of stable storage using stable memory is represented by a single logging site that never fails.

Let $\mathcal{L}$ denote the set of logging sites. Function $L(p)$ denotes the logging site used by process p. Assume that $L(p)$ is a constant function — each process uses a single logging site that does not change. $P(\ell)$, for $\ell \in \mathcal{L}$, denotes the processes that are associated with logging site ℓ. Define functions $L^{\cup}(\mathcal{P})$ and $P^{\cup}(\mathcal{S})$ to be the set-valued domain versions of $L(p)$ and $P(\ell)$ respectively:

$$L^{\cup}(\mathcal{P}) \stackrel{\text{def}}{=} \bigcup_{p \in \mathcal{P}} L(p) \tag{13}$$

$$P^{\cup}(\mathcal{S}) \stackrel{\text{def}}{=} \bigcup_{\ell \in \mathcal{S}} P(\ell) \tag{14}$$

We now repeat, under these new assumptions, the derivation that in Section 4 led us to Property (6), which enforces the always-no-orphans consistency condition.

A process is an orphan process when its state depends on a determinant that is not logged. Let $\mathcal{C}$ denote a set of failed processes and $\mathcal{E}$ denote a set of failed logging sites. Then,

$$p \;\mathit{orphan\ of}\; \mathcal{C}, \mathcal{E} \stackrel{\text{def}}{=} \left(\begin{array}{l} (p \in \mathcal{N} - \mathcal{C}) \;\wedge \\ \exists m : ((p \in \mathit{Depend}(m)) \wedge (\mathit{Log}(m) \subseteq \mathcal{E})) \end{array} \right) \tag{15}$$

A logging site must fail for a process to become an orphan. Negating (15) and quantifying over all p gives the following:

$$\forall m : ((Log(m) \subseteq \mathcal{E}) \Rightarrow (Depend(m) \subseteq \mathcal{C})) \tag{16}$$

Property (16) is necessary and sufficient to guarantee that the failure of $\mathcal{C}$ processes and of $\mathcal{E}$ logging sites creates no orphans. From definitions (13) and (14), it follows that:

$$\mathcal{A} \subseteq \mathcal{B} \quad \Rightarrow \quad P^{\cup}(\mathcal{A}) \subseteq P^{\cup}(\mathcal{B}) \tag{17}$$

$$\mathcal{A} \subseteq \mathcal{B} \quad \Rightarrow \quad L^{\cup}(\mathcal{A}) \subseteq L^{\cup}(\mathcal{B}) \tag{18}$$

where $\mathcal{A}$ and $\mathcal{B}$ are sets of the appropriate type. Noting that $L^{\cup}(P^{\cup}(\mathcal{S})) = \mathcal{S}$, it is straightforward to show that $P^{\cup}(\mathcal{A}) \subseteq P^{\cup}(\mathcal{B}) \Rightarrow \mathcal{A} \subseteq \mathcal{B}$:

⟨⟨Hypothesis assumed⟩⟩
1. $P^{\cup}(\mathcal{A}) \subseteq P^{\cup}(\mathcal{B})$
⟨⟨Modus ponens, 1 and (18)⟩⟩
2. $L^{\cup}(P^{\cup}(\mathcal{A})) \subseteq L^{\cup}(P^{\cup}(\mathcal{B}))$
⟨⟨$L^{\cup}(P^{\cup}(\mathcal{S})) = \mathcal{S}$ and 2⟩⟩
3. $\mathcal{A} \subseteq \mathcal{B}$ $\Box$

Hence, $\mathcal{A} \subseteq \mathcal{B} \equiv P^{\cup}(\mathcal{A}) \Rightarrow P^{\cup}(\mathcal{B})$. We can replace the antecedent of (16) with the equivalent expression $P^{\cup}(Log(m)) \subseteq P^{\cup}(\mathcal{E})$:

$$\forall m : ((P^{\cup}(Log(m)) \subseteq P^{\cup}(\mathcal{E})) \Rightarrow (Depend(m) \subseteq \mathcal{C})) \tag{19}$$

By assumption, $\mathcal{E}$ and $\mathcal{C}$ are constrained: $P^{\cup}(\mathcal{E}) \subseteq \mathcal{C}$. Hence, the following strengthens (19):

$$\forall m : ((P^{\cup}(Log(m)) \subseteq \mathcal{C}) \Rightarrow (Depend(m) \subseteq \mathcal{C})) \tag{20}$$

Universally quantifying (20) over $\mathcal{C}$ we obtain:

$$\forall m : (Depend(m) \subseteq P^{\cup}(Log(m))) \tag{21}$$

Since we want (21) to hold in every state, we obtain the following property:

$$\forall m : \Box(Depend(m) \subseteq P^{\cup}(Log(m))) \tag{22}$$

Property (22), like Property (4) of Section 4, is a safety property. If satisfied, (4) guarantees that no orphans will be created during a run.

If we assume that no more than $f_{\mathcal{E}}$ logging sites can fail at any time, then $\#m$ is stable once it has been logged at more than $f_{\mathcal{E}}$ logging sites. Hence, Property (22) need hold only as long as failures cannot cause $\#m$ to be lost:

$$\forall m : \Box((|Log(m)| \leq f_{\mathcal{E}}) \Rightarrow ((Depend(m) \subseteq P^{\cup}(Log(m))) \tag{23}$$

Finally, we can strengthen Property (23) as in Section 4.2, to obtain the following property, which characterizes causal message logging when there are shared logging sites:

$$\forall m : \Box((|Log(m)| \leq f_{\mathcal{E}}) \Rightarrow \wedge(Depend(m) \subseteq P^{\cup}(Log(m))) \wedge \Diamond(Depend(m) = P^{\cup}(Log(m)))) \tag{24}$$

Using a derivation similar to the one given in Section 6.3, it is straightforward to develop an optimal protocol Π_{ol} that satisfies Property 24. It may be surprising that at this level of refinement, protocol Π_{ol} is essentially identical to protocol Π_{oc} of Section 6.3. The main differences in the protocols is found through further refinement. The most significant difference is in the size of the data structures they to compute $Log(m)$: for Π_{oc}, the data structure is size $O(n^2)$ while for Π_{ol}, the data structure is size $O(|\mathcal{L}|^2)$ where $|\mathcal{L}|$ is the number of logging sites. Details can be found in [Alv96].

References

[AHM93] Lorenzo Alvisi, Bruce Hoppe, and Keith Marzullo. Nonblocking and orphan-free message logging protocols. In *Proceedings of the 23rd Fault-Tolerant Computing Symposium*, pages 145–154, June 1993.

[Alv96] L. Alvisi. *Understanding the Message Logging Paradigm for Masking Process Crashes.* PhD thesis, Cornell University Department of Computer Science, January 1996.

[AM96] L. Alvisi and K. Marzullo. Tradeoffs in implementing optimal message logging protocols. In *Proceedings of the Fifteenth Symposium on Principles of Distributed Computing*, pages 58–67. ACM, June 1996.

[BBG83] Anita Borg, J. Baumbach, and S. Glazer. A message system supporting fault tolerance. In *Proceedings of the Symposium on Operating Systems Principles*, pages 90–99. ACM SIGOPS, October 1983.

[BJ87] Kenneth Birman and Tommy Joseph. Reliable communication in the presence of failures. *ACM Transactions on Computer Systems*, 5(2):47–76, February 1987.

[BSS91] K. Birman, A. Schiper, and P. Stephenson. Lightweight causal and atomic group multicast. *ACM Transactions on Computer System*, 9(3):272–314, August 1991.

[EZ92] E. N. Elnozahy and W. Zwaenepoel. Manetho: Transparent rollback-recovery with low overhead, limited rollback and fast output commit. *IEEE Transactions on Computers*, 41(5):526–531, May 1992.

[EZ94] E.N. Elnozahy and W. Zwaenepoel. On the use and implementation of message logging. In *Digest of Papers: 24 Annual International Symposium on Fault-Tolerant Computing*, pages 298–307. IEEE Computer Society, June 1994.

[Gra77] James N. Gray. Notes on data base operating systems. In R. Bayer, R. M. Graham, and G. Seegmueller, editors, *Operating Systems: An Advanced Course*, pages 393–481. Springer-Verlag, 1977. Lecture Notes on Computer Science 60.

[JV87] T.Y. Juang and S. Venkatesan. Crash recovery with little overhead. In *Proceedings of the 11th International Conference on Distributed Computing Systems*, pages 454–461. IEEE Computer Society, June 1987.

[JZ87] D.B. Johnson and W. Zwaenepoel. Sender-based message logging. In *Digest of Papers: 17 Annual International Symposium on Fault-Tolerant Computing*, pages 14–19. IEEE Computer Society, June 1987.

[JZ90] D.B. Johnson and W. Zwaenepoel. Recovery in distributed systems using optimistic message logging and checkpointing. *Journal of Algorithms*, 11:462–491, 1990.

[Lam78] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, July 1978.

[Pnu77] Amir Pnueli. The temporal logic of programs. In *Proceedings of the Eighteenth Annual Symposium on Foundations of Computer Science*, pages 46–57, November 1977.

[PP83] M.L. Powell and D.L. Presotto. Publishing: A reliable broadcast communication mechanism. In *Proceedings of the Ninth Symposium on Operating System Principles*, pages 100–109. ACM SIGOPS, October 1983.

[RST91] M. Raynal, A. Schiper, and Sam Toueg. The causal ordering abstraction and a simple way to implement it. *Information Processing Letters*, 39(6):343–350, 1991.

[SBY88] R. E. Strom, D. F. Bacon, and S. A. Yemini. Volatile logging in n-fault-tolerant distributed systems. In *Proceedings of the Eighteenth Annual International Symposium on Fault-Tolerant Computing*, pages 44–49, 1988.

[Sch84] Fred B. Schneider. Byzantine generals in action: Implementing fail-stop processors. *ACM Transactions on Computer Systems*, 2(2):145–154, May 1984.

[SS92] A. Sandoz and A. Schiper. A characterization of consistent distributed snapshots using causal order. Technical Report TR92-14, Departement d'Informatique, Ecole Politechnique Fédérale de Lausanne, 1992.

[SW89] A.P. Sistla and J.L. Welch. Efficient distributed recovery using message logging. In *Proceedings of the Eighth Symposium on Principles of Distributed Computing*, pages 223–238. ACM SIGACT/SIGOPS, August 1989.

[SY85] R. B. Strom and S. Yemeni. Optimistic recovery in distributed systems. *ACM Transactions on Computer Systems*, 3(3):204–226, April 1985.

[VJ94] S. Venkatesan and T.Y. Juang. Efficient algorithms for optimistic crash recovery. *Distributed Computing*, 8(2):105–114, June 1994.

Lorenzo Alvisi (On the right, in the light jacket) received his Ph.D. from the Computer Science Department of Cornell University in 1996. He is currently an Assistant Professor in the Department of Computer Sciences at The University of Texas at Austin, where he co-founded the Laboratory for Experimental Software Systems (LESS). Dr. Alvisi's research interests include concurrency, distributed systems and fault tolerance.

Keith Marzullo (On the right, in the light jacket) received his Ph.D. from the Electrical Engineering Department of Stanford University in 1984. He has worked at Xerox Palo Alto, has been on the faculty of Cornell University, and is currently an Associate Professor in the Computer Science and Department of The University of California, San Diego. His research interests are in distributed fault-tolerant systems, both asynchronous and hard-real-time. Professor Marzullo is an Associate Editor for IEEE Transactions on Software Engineering.