RZ 3438 (# 93592) 07/29/02
Computer Science 17 pages

Research Report

Failure Detection Sequencers: Necessary and Sufficient Information about Failures to Solve Predicate Detection

Felix C. Gärtner

Department of Computer Science
Darmstadt University of Technology
D-64283 Darmstadt
Germany
felix@informatik.tu-darmstadt.de

Stefan Pleisch

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland
spl@zurich.ibm.com

Failure Detection Sequencers: Necessary and Sufficient Information about Failures to Solve Predicate Detection

Felix C. Gärtner
Department of Computer Science
Darmstadt University of Technology
D-64283 Darmstadt, Germany
felix@informatik.tu-darmstadt.de

Stefan Pleisch
IBM Research
Zurich Research Laboratory
CH-8803 Rüschlikon, Switzerland
spl@zurich.ibm.com

Abstract

This paper investigates the amount of information about failures needed to solve the predicate detection problem in asynchronous systems with crash failures. In particular, we show that predicate detection cannot be solved with traditional failure detectors, which are only functions of failures. In analogy to the definition of failure detectors, we define a *failure detection sequencer*, which can be regarded as a generalization of a failure detector. More specifically, our failure detection sequencer Σ outputs information about failures *and* about the final state of the crashed process. We show that Σ is necessary and sufficient to solve predicate detection. Moreover, Σ can be implemented in synchronous systems. Finally, we relate sequencers to perfect failure detectors and characterize the amount of knowledge about failures they additionally offer.

1 Introduction

Predicate detection in distributed settings is a well-understood problem and many techniques together with their detection semantics have been proposed [6]. Most of these techniques address predicate detection with the assumption that no faults occur in the system. However, it is desirable to also detect predicates which refer to the operational state of processes, e.g., predicates such as "$x_i = 1 \wedge \mathit{crashed}_i$", where $\mathit{crashed}_i$ is a predicate that is true iff (if and only if) process p_i has crashed. Since $x_i = 1$ might indicate the presence of a lock, the given predicate can be used to formalize special conditions such as "process p_i crashed while holding a lock", which is useful in the context of databases.

In the context of *crash* failures and the *consensus* problem, *failure detectors* have been devised to provide information about failures [3], but they offer "solely" information about failures. To detect general predicates such as the example predicate above, failure detection information needs to be combined with additional information about the internal state of a process. Indeed, while a failure detector may capture the predicate $\mathit{crashed}_i$, it gives no information about the value of x_i.

Ideally, a predicate detection algorithm never erroneously detects a predicate and does not miss any occurrence of the predicate in the underlying computation. As shown in [11], the quality of predicate detection critically depends on the quality of failure detection. This explains why work in [9, 17, 19] puts a restriction on the type of detectable predicates, or [10] weakens the semantics of predicate detection.

In previous work [11], we have investigated predicate detection in an asynchronous system with crash failures and found that it is impossible to solve predicate detection even with a very strong failure detector, the *perfect failure detector* [3]. In this paper, we show that predicate detection cannot be solved with *any* failure detector (as defined in [3]), no matter how strong it is. For example, consider a "real-time perfect" failure detector which makes no mistakes and flags the occurrence of a crash *immediately*. Even this failure detector is insufficient to solve predicate detection. The reason for this impossibility is that failure detectors are only functions of failures. Our proof is a generalization of previous impossibility proofs by the present authors [11] and by Charron-Bost, Guerraoui and Schiper [5]. We attempt to remedy the unpleasant situation caused by the result and (in analogy to the definition of failure detectors) define a *failure detection sequencer*. A failure detection sequencer is a generalization of a failure detector in that it conveys information that is a function of the failures *and* the current history of the system which is under observation. To solve predicate detection, we define a particular failure detection sequencer class Σ, that only gives one additional piece of information: for every crashed process it gives the latest state of the process before this one crashes. We show that Σ is necessary and sufficient to solve predicate detection and consequently is the "weakest failure detection sequencer" to solve predicate detection.

Although Σ is in a sense "stronger" than a perfect failure detector, it is still possible to implement Σ in synchronous systems. Moreover, we show that using Σ it is possible to implement a *synchronizer* for asynchronous crash-affected systems which makes these systems equivalent to purely synchronous systems in terms of the solvability of *time-free* [5] problems. We finally argue that while perfect failure detectors can be viewed as capturing the synchrony of processes, failure detection sequencers in addition also capture the synchrony of communication.

After presenting the system model and defining the problem of predicate detection in Sections 2 and 3, we present our contributions in the following order: First, we show that it

is impossible to achieve predicate detection with any failure detector in the sense of Chandra and Toueg [3] in Section 4. Section 5 introduces the failure detection sequencer abstraction and shows that a particular sequencer Σ is equivalent to predicate detection. In Section 6, we show how to implement Σ and then discuss the strength of Σ in Section 7. Finally, Section 8 concludes the paper.

2 Model

We consider an asynchronous distributed systems in which processes communicate via message passing. This means that no bounds on message transmission time nor on relative process speeds exist. Message delivery is reliable, i.e., a sent message is eventually delivered and no spurious messages are delivered. Processes can fail by crashing, i.e., simply stop to execute steps. Crashed processes do not recover any more. Processes which do not (ever) crash are called *correct*.

2.1 Distributed Computations

A distributed system, called the *application system*, consists of a finite set Π of n processes $p_1, p_2, \ldots, p_n$ (called *application processes*). Each process p_i has a local state s_i (defined by the values assigned to its local variables) and performs atomic state transitions according to a local algorithm A. Such a state transition is also called an *event*. Sending and receiving a message also results in a state change. If a process p_i sends a message in state s_i which is received by process p_j resulting in state s_j, we say that s_i and s_j *correspond*.

We define a relation of *potential causality* (denoted "$\rightarrow$") [2] on local states as the transitive closure of the following two relations:

- $s \rightarrow s'$ if s and s' happen on the same process and s happens before s'.
- $s \rightarrow s'$ if s and s' happen on different processes and s' and s correspond.

A *local history of* p_i is an (infinite) sequence $s_1, s_2, \ldots$ of states. A *distributed computation* is defined as a set of local histories, one for every process. A *global state* of the computation is a vector $G = (s_1, s_2, \ldots, s_n)$ of local states, one for each process. Each local state identifies a point in the local history of a process and thus is equivalent to the set of all local states the process went through to reach its "current" local state. A global state G is *consistent* if the union of these sets (of all local states in G) is left-closed with respect to $\rightarrow$, i.e., if a state s is in this set and $s' \rightarrow s$, then s' must also be in this set. The set of all global states of a computation together with $\rightarrow$ define a lattice [15].

We assume the existence of a discrete global clock. Processes do not have access to this global clock; it is merely a fictionous device to simplify presentation. Let $\mathcal{T}$ denote the range of output values of the global clock. For simplicity we think of $\mathcal{T}$ to be the set of natural numbers.

2.2 Failure Detectors

A *failure detector* is a device that can be queried at any time $t \in \mathcal{T}$ and outputs the set of processes that it suspects to have crashed at time t.

We adopt the formal definitions of failure detectors by Chandra and Toueg [3]. A *failure pattern* F is a mapping from $\mathcal{T}$ to the powerset of Π. The value of $F(t)$ specifies the set of application processes that have crashed until time $t \in \mathcal{T}$. A *failure detector history* H is a mapping from $\Pi \times \mathcal{T}$ to the powerset of Π. The value of $H(p,t)$ denotes the return value of the failure detector module for process p at time t, i.e., if p queries the failure detector at time t, $H(m,t)$ contains the set of processes suspected at that time.

A *failure detector* $\mathcal{D}$ maps a failure pattern F to a set of failure detector histories. The set of histories returned by the failure detector satisfy certain accuracy and completeness properties. A perfect failure detector satisfies *strong accuracy* and *strong completeness*:

- Strong accuracy: no process is suspected before it crashes. Formally:

$$\forall F.\forall H \in \mathcal{D}(F).\forall t \in \mathcal{T}.\forall p,q \in \Pi \setminus F(t).p \notin H(q,t)$$

- Strong completeness: a crashed process is eventually permanently suspected by every correct process. Formally:

$$\forall F.\forall H \in \mathcal{D}(F).\forall p \in \Pi.\forall q \in \mathit{correct}(\Pi).\forall t \in \mathcal{T}.p \notin F(t) \wedge p \in F(t+1) \Rightarrow \exists t'.\forall t'' > t'.p \in H(q,t')$$

The set of all perfect failure detectors is denoted by $\mathcal{P}$. In the following, we will sometimes use the symbol $\mathcal{P}$ as a shorthand for any failure detector from $\mathcal{P}$.

2.3 Runs and Steps

Chandra and Toueg [3] define a computation (which they call a *run*) to be a tuple $R = (F, \mathcal{D}, I, S, T)$, where S is a sequence of algorithm steps and T is a sequence of increasing time values when these steps are taken. Steps are defined with respect to an algorithm which in turn is a collection of deterministic automata. We define a run in a slightly different but equivalent manner. Instead of S and T we use two functions: a *step function* S_s from $\mathcal{T}$ to the set of all algorithm steps, and a process function S_p from $\mathcal{T}$ to Π. Briefly spoken, $S_p(t)$ denotes the process which takes a step at time t and $S_s(t)$ identifies the step which was taken. Without loss of generality, we assume that at any instance of time at most one process takes a step. If no process takes a step at time t, both functions evaluate to $\perp$. A computation then is a tuple $R = (F, \mathcal{D}, I, S_s, S_p)$.

In predicate detection, which is defined in the following section, we wish to detect whether a predicate holds on the state of processes. We assume that the state resulting from an algorithm step contains "enough information" for predicate detection purposes, e.g., if we are interested in detecting whether or not a process has reached line x, a "program counter" must be included in the local state of a process. The most recent step therefore can be used to infer the state which the process is in after executing that step. In this paper, we use the terms *state* and *step* interchangeably.

3 Predicate Detection

To detect predicates in the application system (see Section 2.1), the application system is extended with a set Φ of m *monitor processes* $b_1, \ldots, b_m$. The sets Π and Φ together form the *observation system*.

While application processes may crash, we assume, for simplicity, that monitor processes do not. Crashes of application processes do not change the local state of the process. However, the operational state of a process p_i is modeled by a virtual boolean variable $crashed_i$ on every monitor. The global state of the system together with the vector of *crashed* variables defines the *extended global state* of the system.

The task of the monitor processes is to observe the application processes and invoke a special primitive *detected* if the state of the system satisfies a certain predicate. A predicate ϕ is a boolean function on the extended global state of the application system. For example, the predicate $x_i = 2 \wedge crashed_i$ is true in a global state if the variable x_i of p_i equals 2 and p_i has crashed. We say that ϕ *holds* in a computation c iff there exists a consistent global state in c such that ϕ is true in that state.

In our version of predicate detection, monitors can observe multiple predicates simultaneously. More specifically, the predicate detection algorithm maintains a set S of currently active predicates. A special primitive $fork(\phi)$ can be used to add a predicate ϕ to this set. Whenever some $\phi \in S$ is found to hold in the computation, the predicate detection algorithm indicates this by pointing to ϕ, i.e., by calling $detected(\phi)$. Formally, detecting any $\phi \in S$ corresponds to detecting the disjunction of all such ϕ. This formulation of predicate detection has the important advantage of allowing us to increase the set of observed predicates at runtime. In other words, it does not matter when a predicate ϕ is added to S. Even if ϕ held "early" in the computation and $fork(\phi)$ is invoked very late (e.g., after hours), then still the algorithm must eventually invoke $detected(\phi)$ [1]. In this sense, our predicate detection concept is *adaptive* and thus slightly more general than other definitions of predicate detection (e.g., *perfect predicate detection* [11]). This is reflected in the following definition:

Definition 1 (predicate detection) *A predicate detection algorithm is a distributed algorithm running on the observation system with an input operation fork() and an output operation detected(). Using fork(ϕ) a new predicate can be added to an initially empty set S of predicates. The algorithm must satisfy the following properties:*

- *(Safety) If a monitor invokes detected(ϕ) then ϕ holds in the computation and $\phi \in S$.*
- *(Liveness) If $\phi \in S$ and ϕ holds in the computation, then eventually a monitor must invoke detected(ϕ).*

Our definition of predicate detection makes no reference to a specific implementation. Generally, one expects application processes to use causal broadcast [2] to consistently disseminate information about every local state change to all monitor processes. But this is not demanded by the specification. Furthermore, there is no indication how monitors keep track of the changes of *crashed* values of processes, i.e., we do not postulate the existence of a special type of failure detector in the specification. However, failure detection can be considered a special case of predicate detection on the extended state space where the predicate to be detected consists only of the *crashed* variables of processes. This highlights the close relationship between failure detection and predicate detection which is studied in the following sections.

Note that the meaning of "ϕ holds in the computation" corresponds to the detection modality $possibly(\phi)$ [7, 14]. Detecting $possibly(\phi)$ involves constructing the entire computation lattice in the general case. The lattice represents all possible observations; hence, an

[1] Later in Section 5.2 we show how this property can be implemented in asynchronous systems.

observation is a path through the lattice. For simplicity we restrict our attention to *observer-independent predicates* [4]. For these types of predicates it is sufficient to construct a single observation, i.e., a single path through the lattice, to check whether ϕ holds in the observation. For example, stable predicates are observer-independent (a predicate is stable iff once it holds it holds forever). However, not all observer independent predicates are stable. For example, predicates which are local to a single process are observer-independent but may not be stable.

4 Impossibility of Predicate Detection in a Faulty Environment using Failure Detectors

In this section we show that predicate detection cannot be solved with any failure detector in the sense of Chandra and Toueg [3]. This is because failure detectors are "functions of failures", i.e., the failure detector $\mathcal{D}$ is a function which maps a failure pattern F to some element of an arbitrary range $\mathcal{G}$. The proof is based on the assumption that apart from using failure detectors and (asynchronous) messages, no information can flow between processes. Messages sent by application processes to monitor processes for the sake of predicate detection are called *control messages*. The impossibility holds even if we assume that state changes on the application processes and the broadcast of control messages happen atomically.

The problem underlying this impossibility is to know whether there are still messages in transit from process p_i even if the failure detector already suspects p_i. Indeed, if we assume that an application process broadcasts a message with every local state change, the latest state of p_i is reflected by the "latest" control messages received by the monitor m. Using causal broadcast [2] and the information from these control messages m can construct the latest state of p_i. Crash detections, however, arrive out of sequence with the process events (and also the control messages); they are obtained by querying a local failure detector module. When m detects the failure of p_i, control messages from p_i may still be in transmission from p_i to m (see Figure 1 (a)). Hence, m needs to wait for the arrival of these messages before considering the failure of p_i. However, in an asynchronous system, these messages may take a long time to arrive at m; m cannot distinguish at a given point in time, whether a message is still in transit and has been delayed (see Figure 1 (a)) or whether no message will arrive from p_i any more [8] (see Figure 1 (b)).

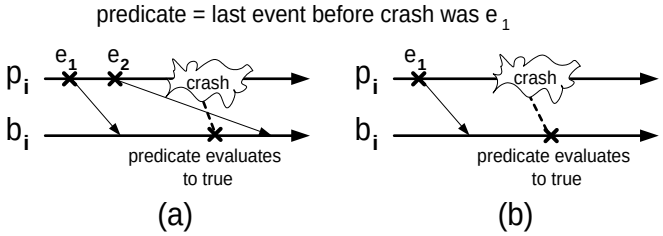

Figure 1: Intuitive reason for the impossibility result of Theorem 1.

Theorem 1 *It is impossible to solve predicate detection with any failure detector $\mathcal{D}$.*

For lack of space, the formal proof is relegated to Appendix A.1. In the next section we consider a way of circumventing the impossibility by extending the concept of a failure detector to a component that is also useful in the context of predicate detection. We call such a component *failure detection sequencer*.

5 Failure Detection Sequencers

The failure detector abstraction was introduced by Chandra and Toueg [3] to characterize different system models with respect to the solvability of problems in fault-tolerant computing.

We take a similar approach as [3] and devise an oracle that encodes enough information to solve predicate detection in asynchronous systems with process crashes. As shown in the previous section, information about failures alone is not sufficient. Hence, our oracle also needs to provide information about the state of the process when it crashed.

5.1 Definition

We now define a *failure detection sequencer* Σ, which consists of a set of passive modules, one for each monitor process. The sequencer can be queried by the monitor and returns an array of size n. The value at index i of the array is either $\perp$ or contains a predicate ϕ on the local state of process p_i. Informally spoken, the latter means that p_i has crashed and that its final state satisfied ϕ. The predicate ϕ may have different forms, e.g., indicate a unique sequence number of the step last performed by p_i. Let $\mathcal{A}$ denote the set of all possible array values, i.e., combinations of $\perp$ and local predicates, which can be returned by Σ. Formally, Σ is defined as follows:

A *sequencer history* H_Σ is a mapping from $\Phi \times \mathcal{T}$ to $\mathcal{A}$. The value of $H_\Sigma(m,t)$ indicates the return value of Σ at monitor m if it is queried at time t. If $H_\Sigma(m,t)[i] = s$, then m suspects p_i at time t to be in s ($s \neq \perp$). A *failure detection sequencer* Σ maps a failure pattern F, a step function S_s and a process function S_p to a set of sequencer histories.

Given a time t, the most recent step of a process p_i can be determined by inspecting S_s and S_p. If p_i has not executed any step, then the most recent step is denoted by ϵ. Formally, the *most recent step of* p_i *at* t given S_s and S_p is s iff

$$most\_recent\_step(p_i, t, S_s, S_p) : \exists t' \leq t.(S_s(t') = s) \wedge (S_p(t') = p_i) \wedge (\forall t''.t' < t'' < t.S_p(t'') \neq p_i)$$

We require that the set of all possible sequencer histories H_Σ satisfies the following two properties:

- (Accuracy) No process is incorrectly suspected to be in state s. Formally:

$$\forall m.\forall p_i.\forall t.H_\Sigma(m,t)[i] = s \neq \perp \Rightarrow p_i \in F(t) \wedge (s = most\_recent\_step(p_i, t, S_s, S_p))$$

- (Completeness) If p crashes, then eventually every monitor will permanently suspect p to be in some state. Formally:

$$\forall m.\forall p_i.\forall t.p \in F(t) \Rightarrow \exists t' \geq t.\forall t'' \geq t'.H_\Sigma(m,t'')[i] \neq \perp$$

Since the accuray requirement has a conjunction in the consequent, it is possible to separate it into a step accuracy part and a crash accuracy part. Crash accuracy corresponds to

strong accuracy of Chandra and Toueg [3] ("no process is suspected before it crashes"), while step accuracy would mean that a non-$\perp$ sequencer output for process p_i at time t always equals the state which p_i is in *at the same moment* (i.e., at time t). Clearly, this property has only trivial solutions (i.e., a solution which always outputs $\perp$) since asynchronous message passing does not allow instantaneous message delivery. However, the combination of step accuracy and crash accuracy makes sense again since crashes "freeze" the state of a process so that there is no danger of state change once the sequencer has suspected that process.

We have called the new device a "sequencer" because it allows to implement causal order on failure detection events, as we now explain. Using Σ it is possible to infer the crashed state of a process at the moment it is suspected. This means that it is possible to know how many control messages are in transit. Hence, the "delivery" of the suspicion can be delayed until all causally preceding events have been delivered; Σ can be used to "sequence" crash notifications, as shown in the following section.

5.2 Equivalence to Predicate Detection

Now we investigate the power of failure detection sequencers and show that they are sufficient and necessary to solve predicate detection. First we consider sufficiency.

The idea of implementing predicate detection using Σ is to embed crash events consistently into the causal order $\rightarrow$ of events in a computation. For this purpose, the algorithm shown in Figure 2 uses *causal broadcast* [2] (using primitives *c-bcast* and *c-deliver*) to disseminate information about state changes to all monitors and to withhold issuing the crash occurrence when Σ suspects p_i after some state s until the state of p_i has indeed reached state s. This is done using a vector *def_crash*$[i]$ (for "deferred crash").

The adaptability of predicate detection is implemented by using a variable *history*, a sequence of global states. Whenever a new predicate ϕ is issued using the *fork* command, the entire history is checked whether or not ϕ held in the past. For lack of space the proof of the following theorem is given in Appendix A.2.

Theorem 2 *Predicate detection can be solved using* Σ.

We now show that Σ is necessary to solve predicate detection. To do this we assume the existence of an abstract algorithm *PD* that solves predicate detection on a given computation. Then we give an algorithm that emulates the output vector of Σ using *PD*.

Similar to the predicate detection algorithm in Figure 2 we instruct application processes to send a control message to all monitors if a local event happens. These control messages are used to fork an increasing number of instances of *PD*. Initially, a single instance for the predicate "p_i crashed in initial state" is started for every process p_i. When the first control message (i, s) arrives, a new instance is *fork*ed for the predicate "p_i crashed in state s". This is done whenever a new control message arrives.

The *output* vector which simulates the output of Σ is initialized with $\perp$ values and only changed, if one of the instances of predicate detection terminates by issuing *detected*(ϕ). This indicates that a process crashed in some state. The algorithm reflects this by changing the corresponding entry in *output*. The change is permanent since the state in which a process crashes does not change anymore. The formal proof of the following theorem is given in Appendix A.3.

Theorem 3 *If predicate detection is solvable, then* Σ *can be implemented.*

```
1  On every application process p_i:
2    ⟨whenever a state change from s to s' happens⟩ do
3      c-bcast (i, s) to all monitors
4  On every monitor process m_j:
5    variables:
6      state[1..n] of ⟨local state information⟩ init ⟨initial states of processes⟩
7      crashed[1..n] of boolean init false
8      def_crashed[1..n] of {⊥} ∪ ⟨local state information⟩
9      history sequence of ⟨(state, crashed)⟩ init ⟨initial state⟩
10     S set of ⟨global predicates⟩ init ∅
11   algorithm:
12     do forever
13       case ⟨next event⟩ of            {* three cases possible *}
14       case 1: ⟨(i, s) is c-delivered⟩
15         state[i] := s
16         history := history · (state, crashed)
17         if ∃φ ∈ S.φ(state, crashed) then detected(φ) endif
18         if def_crash[i] = state[i] then crashed[i] := true endif
19         history := history · (state, crashed)
20         if ∃φ ∈ S.φ(state, crashed) then detected(φ) endif
21       case 2: ⟨Σ suspects p_i in s⟩
22         if state[i] = s then
23           crashed[i] := true
24           history := history · (state, crashed)
25           if ∃φ ∈ S.φ(state, crashed) then detected(φ) endif
26         else {* state[i] ≠ s *}
27           def_crash[i] := s
28         endif
29       case 3: ⟨fork(φ) is called⟩
30         S := S ∪ {φ}
31         if ∃s_i ∈ history.φ(s_i) then detected(φ) endif
32       end {* case *}
33     end {* do forever *}
```

Figure 2: Solving predicate detection using Σ. The primitives *c-bcast* and *c-deliver* denote causal broadcast and causal message delivery, respectively. The operator $\cdot$ denotes concatenation of sequences. Furthermore, the choice of the case statement is supposed to happen in a fair manner (e.g., event handling is performed using first-come first-serve).

It is interesting to study the role of adaptiveness (i.e., the ability to "restart" predicate detection via *fork*) in the proof of Theorem 3. To see this, consider a definition of predicate detection without adaptiveness, i.e., it is merely possible to start instances of *PD* at the beginning of the computation. Not knowing the way in which the computation will proceed, it is necessary to invoke an instance of predicate detection for *every* state a process may reach. Hence, non-adaptive predicate detection can be used to implement Σ as long as the state space of a process is finite. Adaptiveness allows to invoke instances of predicate detection "on demand". This means that — given infinite state space — while there is no bound on the number of calls to *fork*, the number of "parallel" instances of predicate detection is always finite.

The following theorem is an immediate consequence of Theorems 2 and 3. It can be rephrased as showing that Σ is the "weakest failure detector" for solving predicate detection. The quotation marks are important, because from Theorem 1 we know that we should not

```
1  On every application process p_i:
2    ⟨whenever a state change (s, s') happens⟩ do
3      c-bcast (i, s) to all monitors
4  On every monitor process m_j:
5    variables:
6      output[1..n] of {⊥} ∪ ⟨process state information⟩ initially ⊥
7    algorithm:
8      for all i ∈ {1, ..., n} do begin
9        fork("p_i crashed in initial state") end
10     do forever
11       ⟨wait until (i, s) is c-delivered or detected(φ) is invoked⟩
12       if ⟨(i, s) was c-delivered⟩ then
13         fork("p_i crashed in state s")
14       elsif ⟨detected(φ) was called⟩then
15         {* φ is "p_i crashed in s" *}
16         output[i] := s
17       endif
18     end {* do forever *}
```

Figure 3: Emulating Σ using a predicate detection algorithm. State changes and sending control messages on application processes is assumed to happen atomically. Event handling in line 11 is again performed in a fair manner, e.g., using first-come first-serve.

call Σ a failure detector.

Theorem 4 *Solving predicate detection is equivalent to implementing* Σ*.*

6 Implementing Σ

The sequencer Σ is a rather strong device and its strength makes it a highly desirable tool in crash-affected systems. Hence, the question naturally arises on how to implement Σ in "real" systems. First, consider *synchronous systems*, i.e., systems where bounds on message delivery delays and relative processing speeds exist. In synchronous systems, Σ can easily be implemented, for instance, by the algorithm in Figure 4. This algorithm is an adaption of the algorithm for implementing a perfect failure detector in synchronous systems presented by Tel [16]. With every local step, process p_i decrements a special timer variable r, one for every remote process. Upon message reception from process p_j ($j \neq i$), the timer is reset to the inital value δ, which is computed from the maximum message delivery delay and the maximum processing speed difference. If p_i fails to receive a message from p_j before the timer elapses, then p_j is suspected by p_i.

To see that the algorithm indeed implements Σ, we need to show that it satisfies the accuracy and completeness properties given in Section 5.1. The proof of the completeness property is the same as for perfect failure detectors. To see that the accuracy property is satisfied, consider the sequence of "alive" messages received by Σ. As these messages are sent and arrive in FIFO order[2], the failure detector also receives the correct sequence of state information. If p_j crashes, the final message received by p_i ($i \neq j$) is also the final message

[2]FIFO broadcasts are implementable in synchronous systems, as they can even be implemented in asynchronous systems [12].

```
1  On every process p_j:
2    with ⟨every step⟩ FIFOsend "alive in state s" to all
3  On every process p_i:
4    variables:
5      D_i[1..n] init (⊥, ..., ⊥) {* sequencer output *}
6      r_i[1..n] init (δ, ..., δ) {* timers *}
7      S_i[1..n] init ⟨initial states of p_1, ..., p_n⟩
8    algorithm:
9      upon FIFOreceive "alive in state s" from p_j do
10       ⟨reset timer r_i[j] to δ⟩
11       S_i[j] := s
12     upon ⟨expiry of timer r_i[j]⟩ do
13       D_i[j] := S_i[j]
```

Figure 4: Implementing Σ in synchronous systems. The value δ is a local timeout value computed from the global boundary on message delivery delay and relative processing speeds.

which was sent by p_j. This implies that the state information given in that message is a true indication of the most recent step performed by p_j.

Theorem 5 *In a synchronous system the output of the algorithm in Figure 4 satisfies the accuracy and completeness conditions of* Σ.

Now consider a system without bounds on relative process speeds but bounded communication delays (i.e., asynchronous processes with *synchronous* communication). In such systems, Σ is implementable if any $\mathcal{D} \in \mathcal{P}$ is given. The algorithm is shown in Figure 5 and is similar to the one in Figure 4. Here the timing bound δ refers to the synchrony of the communication channels. Completeness is achieved through the completeness of $\mathcal{D}$ and the fact that the timer eventually runs out. Accuracy is satisfied because of the accuracy of $\mathcal{D}$, the FIFO property of messages (as above), and the fact that after expiry of the timer, no message can be in transit (bounded communication delays).

```
8    algorithm:
9      upon FIFOreceive "alive in state s" from p_j do
10       S_i[j] := s
11     upon ⟨D suspects p_j⟩ do
12       ⟨reset timer r_i[j] to δ⟩
13     upon ⟨expiry of timer r_i[j]⟩ do
14       if⟨D suspects p_j⟩ then
15         D_i[j] := S_i[j]
16       endif
17
```

Figure 5: Implementing Σ using $\mathcal{D} \in \mathcal{P}$ and synchronous communication (lines 1 to 7 are the same as in Figure 4). The value δ is a local timeout value computed from the global boundary on message delivery delay.

Theorem 6 *In a system with asynchronous processes, synchronous communication, and any $\mathcal{D} \in \mathcal{P}$, the output of the algorithm in Figure 5 satisfies the accuracy and completeness conditions of Σ.*

We discuss the relationship between perfect failure detectors and Σ in more detail in the following section.

7 Discussion

We have shown that predicate detection cannot be solved with a perfect failure detector. However, it is solvable using failure detection sequencer Σ. In a sense this means that Σ is "stronger" than a perfect failure detector. Since both abstractions can be implemented in synchronous systems, a perfect failure detector seems to "loose" some information at its interface which a sequencer retains. In this context, two questions arise which we now discuss: (1) How can this difference in information be characterized, and (2) how much information (if any) does a sequencer loose compared to a fully synchronous system?

Regarding question (1), it seems that the synchrony of communication is the aspect which Σ (in contrast to perfect failure detectors) encapsulates. Consider for example an additional oracle Δ which can be queried whether or not the communication channel to a process p_j is empty. Both oracles, Δ and any $\mathcal{D} \in \mathcal{P}$, are incomparable, since they cannot emulate eachother in asynchronous crash-affected systems. However, using Δ instead of the timeout mechanism in the algorithm of Figure 5 yields Σ. Hence, knowledge about the synchrony of communication channels is all that is needed to strengthen a perfect failure detector to Σ. Conversely, this information can be regarded as being "lost" at the interface of a perfect failure detector.

Regarding question (2), we now argue that Σ retains the "full" information present in synchronous systems. Using Σ, it is possible to implement a *synchronizer* [1] for asynchronous crash-affected systems. A synchronizer is a distributed algorithm that allows asynchronous processes to proceed in *rounds*. For this, the synchronizer generates a sequence of "clock-pulses" at each process which separate the rounds. With every pulse, a process is allowed to send at most one message to one of its neighbors. The synchronizer ensures that all messages sent at the beginning of round r are received within round r. It also ensures that every correct process (i.e., a process that does not fail) participates in infinitely many rounds.

Since the failure detection sequencer makes it possible to identify the "final" message from a crashed process, it is possible to implement such a synchronizer just like in the fault-free case [16, p. 408]: At the beginning of round r, every surviving process sends exactly one message m_r to every other process (using reliable broadcast [12]). The "application message" which the process might send in round r is packed together with this synchronizer message to form a single message. A process p_i waits until, for every other process p_j, either (a) m_r is received or (b) Σ suspects p_j. Note that in the latter case it is possible to distinguish the two cases where p_j crashed *before* or *after* sending the message m_r. (This distinction is *not* possible with a perfect failure detector.) Waiting for m_r is important in order to satisfy the specification of the synchronizer, as no other way exists to prevent application messages from round r to be received in round $r + 1$ or later.

The pulses generated by the synchronizer resemble a form of global logical time. Such a time is present in synchronous systems and so the synchronizer transforms the asynchronous system into a synchronous system, with the exception of global real time. In other words,

time-free applications [5] perceive an asynchronous system agmented with Σ as equally strong as a synchronous system. Hence, Σ can be regarded as a form of failure detector which offers applications "full" synchrony without referring to a global clock.

8 Future Work

Many open issues for future work remain: For instance, can other protocols (like those used for solving *consensus*) exploit the additional power of failure detection sequencers to improve efficiency? Another interesting issue is whether other (possibly weaker) classes of failure detection sequencers are meaningful in asynchronous systems and offer more information than failure detectors. An obvious candidate would be an "eventually accurate" failure detection sequencer $\diamond\Sigma$. However, we conjecture that $\diamond\Sigma$ is equivalent to $\mathcal{P}$ with respect to the problems it allows to solve.

Acknowledgments

We wish to thank Rachid Guerraoui for his comments on an earlier version of this paper. The first author wishes to thank Ted Herman for many helpful discussions on the topic of failure detection.

References

[1] Baruch Awerbuch. Complexity of network synchronization. *Journal of the ACM*, 32(4):804–823, October 1985.

[2] K.P. Birman and T.A. Joseph. Reliable communication in the presence of failures. *ACM Transactions on Computer Systems*, 5(1):47–76, February 1995.

[3] Tushar Deepak Chandra and Sam Toueg. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM*, 43(2):225–267, March 1996.

[4] Bernadette Charron-Bost, Carole Delporte-Gallet, and Hugues Fauconnier. Local and temporal predicates in distributed systems. *ACM Transactions on Programming Languages and Systems*, 17(1):157–179, January 1995.

[5] Bernadette Charron-Bost, Rachid Guerraoui, and André Schiper. Synchronous system and perfect failure detector: Solvability and efficiency issues. In *International Conference on Dependable Systems and Networks (IEEE Computer Society)*, 2000.

[6] Craig M. Chase and Vijay K. Garg. Detection of global predicates: Techniques and their limitations. *Distributed Computing*, 11(4):191–201, 1998.

[7] Robert Cooper and Keith Marzullo. Consistent detection of global predicates. *ACM SIGPLAN Notices*, 26(12):167–174, December 1991.

[8] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, April 1985.

[9] Vijay K. Garg and J. Roger Mitchell. Distributed predicate detection in a faulty environment. In *Proceedings of the 18th IEEE International Conference on Distributed Computing Systems (ICDCS98)*, 1998.

[10] Felix C. Gärtner and Sven Kloppenburg. Consistent detection of global predicates under a weak fault assumption. In *Proceedings of the 19th IEEE Symposium on Reliable Distributed Systems (SRDS2000)*, pages 94–103, Nürnberg, Germany, October 2000. IEEE Computer Society Press.

[11] Felix C. Gärtner and Stefan Pleisch. (Im)Possibilities of predicate detection in crash-affected systems. In *Proceedings of the 5th Workshop on Self-Stabilizing Systems (WSS 2001)*, number 2194 in Lecture Notes in Computer Science, pages 98–113, Lisbon, Portugal, October 2001. Springer-Verlag.

[12] Vassos Hadzilacos and Sam Toueg. A modular approach to fault-tolerant broadcasts and related problems. Technical Report TR94-1425, Cornell University, Computer Science Department, May 1994.

[13] Leslie Lamport. How to write a proof. *American Mathematical Monthly*, 102(7):600–608, August/September 1995.

[14] Keith Marzullo and Gil Neiger. Detection of global state predicates. In *Proceedings of the 5th International Workshop on Distributed Algorithms (WDAG91)*, pages 254–272, 1991.

[15] Friedemann Mattern. Virtual time and global states of distributed systems. In M. Cosnard et al., editor, *Proceedings of the International Workshop on Parallel and Distributed Algorithms*, pages 215–226, Chateau de Bonas, France, 1989. Elsevier Science Publishers. Reprinted on pages 123–133 in [18].

[16] Gerard Tel. *Introduction to Distributed Algorithms*. Cambridge University Press, second edition, 2000.

[17] Subbarayan Venkatesan. Reliable protocols for distributed termination detection. *IEEE Transactions on Reliability*, 38(1):103–110, April 1989.

[18] Zhonghua Yang and T. Anthony Marsland, editors. *Global States and Time in Distributed Systems*. IEEE Computer Society Press, 1994.

[19] Pei yu Li and Bruce McMillin. Fault-tolerant distributed deadlock detection/resolution. In *Proceedings of the 17th Annual International Computer Software and Applications Conference (COMPSAC'93)*, pages 224–230, November 1993.

A Proofs

We now give the formal proofs of the theorems. Proofs are written in a structured style similar to proof trees of interactive theorem proving environments. This approach is advocated by Lamport who promises that this style "makes it much harder to prove things that are not true" [13]. The proof is a sequence of numbered proof steps at different levels. Every proof step has a proof which may be refined at lower levels by additional proof steps. For example,

proof step $\langle 1 \rangle 2$ is the second proof step on level 1. Proofs may also be read in a structured way, for example, by reading only the top level proof steps and going into sublevels only when necessary.

A.1 Proof of Theorem 1

ASSUME: There exists an algorithm A which solves predicate detection using $\mathcal{D}$.
PROVE: False

$\langle 1 \rangle 1$. Consider a run $R_1 = (F, \mathcal{D}(F), I, S_s, S_p)$ in which p crashes without executing a single step. Consider the predicate $\phi \equiv$"p crashed in initial state". Eventually (say at time t_1), A will detect ϕ.
PROOF: Follows from the liveness property of predicate detection and the assumption that A solves predicate detection. □

$\langle 1 \rangle 2$. Consider a run R_2 with the same failure pattern F, but different S_s and S_p, where p executes a step s_1 before it crashes. Algorithm A never detects ϕ.
PROOF: From the safety property of predicate detection and the assumption that A solves predicate detection. □

$\langle 1 \rangle 3$. In run R_2, eventually (say at time t_2) A receives a control message which includes information by p about executing step s_1.
PROOF: Follows from step $\langle 1 \rangle 2$, the assumption that no other means exist to communicate local state changes, the atomicity of step execution and control message sending on p, and the reliable channel assumption. □

$\langle 1 \rangle 4$. $t_2 \leq t_1$
  $\langle 2 \rangle 1$. ASSUME: $t_2 > t_1$
    PROVE: False
    $\langle 3 \rangle 1$. In run R_2, A does not detect ϕ at t_1.
      PROOF: Follows from step $\langle 1 \rangle 2$ (A never detects ϕ in R_2). □
    $\langle 3 \rangle 2$. $\mathcal{D}(F)$ is the same in both runs R_1 and R_2.
      PROOF: Follows from the fact that F is the same and that $\mathcal{D}$ is a function of F. □
    $\langle 3 \rangle 3$. In run R_1, A does not detect ϕ at t_1.
      From steps $\langle 3 \rangle 1$, $\langle 3 \rangle 2$ and the fact that no other communication occurs between p and m. □
    $\langle 3 \rangle 4$. Q.E.D.
      PROOF: Step $\langle 3 \rangle 3$ contradicts step $\langle 1 \rangle 1$. □
  $\langle 2 \rangle 2$. Q.E.D.
    PROOF: Follows indirectly from step $\langle 2 \rangle 1$. □

$\langle 1 \rangle 5$. Q.E.D.
  PROOF: Step $\langle 1 \rangle 4$ violates the asynchrony condition of communication. □

A.2 Proof of Theorem 2

ASSUME: Σ is available.
PROVE: Predicate detection can be solved.

$\langle 1 \rangle 1$. The variable S implements the specification variable of the same name.
  PROOF: The specification variable collects the predicates which have been given to *fork*. This is done by the algorithm in Figure 2 at lines 29 and 30. Hence, S can be regarded as a true implementation of the specification variable. □

⟨1⟩2. The tuple $(state, crashed)$ always contains a consistent global state on the extended state space of the computation.

PROOF: Follows from the use of the *c-bcast* and *c-deliver* primitives, the definition of causal order, the way in which *def_crash* is used, and the accuracy property of Σ. □

⟨1⟩3. Every state in *history* is a consistent global state on the extended state space of the computation.

PROOF: Follows from step ⟨1⟩2 and the way in which new states are appended to *history*. □

⟨1⟩4. If ϕ holds in the computation, then eventually monitor m_j will construct a global state $G = (state, crashed)$ on the extended state space such that ϕ holds in G.

PROOF: Follows from the use of the *c-bcast* and *c-deliver* primitives (their liveness), the fact that ϕ is observer-independent, the completeness property of Σ and the fair event scheduling assumption. □

⟨1⟩5. If ϕ holds in the computation, then eventually m_j will add a global state G to *history* such that ϕ holds in G.

PROOF: Follows from step ⟨1⟩4 and the algorithm. □

⟨1⟩6. The algorithm in Figure 2 satisfies the Safety property of predicate detection.

ASSUME: The monitor invokes $detected(\phi)$.

PROVE: ϕ holds in the computation and $\phi \in S$.

⟨2⟩1. ASSUME: $detected(\phi)$ is invoked in lines 17, 20, or 25.

PROVE: Q.E.D.

⟨3⟩1. $\phi \in S$ and S reflects the specification variable of the same name.

PROOF: Follows from the algorithm and step ⟨1⟩1. □

⟨3⟩2. The monitor has constructed a consistent global state $G = (state, crashed)$ on the extended state space such that ϕ holds in G.

PROOF: Follows from the algorithm. □

⟨3⟩3. There exists a consistent global state of the computation such that ϕ holds in that state.

PROOF: Follows from steps ⟨3⟩2 and ⟨1⟩2. □

⟨3⟩4. Q.E.D.

PROOF: From step ⟨3⟩3 follows that ϕ holds in the computation and from step ⟨3⟩1 follows that $\phi \in S$. □

⟨2⟩2. ASSUME: $detected(\phi)$ is invoked in line 31.

PROVE: Q.E.D.

⟨3⟩1. $\phi \in S$

PROOF: Follows from the algorithm and step ⟨1⟩1. □

⟨3⟩2. There exists a global state G in *history* such that ϕ holds in G.

PROOF: Follows from the algorithm. □

⟨3⟩3. There exists a state in the computation such that ϕ holds in that state.

PROOF: Follows from steps ⟨3⟩2 and ⟨1⟩3. □

⟨3⟩4. Q.E.D.

PROOF: From step ⟨3⟩3 follows that ϕ holds in the computation and from step ⟨3⟩1 follows that $\phi \in S$. □

⟨2⟩3. Q.E.D.

PROOF: Steps ⟨2⟩1 and ⟨2⟩2 cover all cases. □

⟨1⟩7. The algorithm in Figure 2 satisfies the Liveness property of predicate detection.

ASSUME: ϕ holds in the computation and $\phi \in S$.

PROVE: Eventually the monitor invokes $detected(\phi)$.

⟨2⟩1. Eventually (at time t_1), the monitor constructs a global state $G = (state, crashed)$ on the extended state space such that ϕ holds in G.
PROOF: Follows from step ⟨1⟩4. □
⟨2⟩2. At some point t_2 in time $fork(\phi)$ was invoked.
PROOF: Follows from the assumption that $\phi \in S$ and step ⟨1⟩1. □
⟨2⟩3. ASSUME: $t_1 < t_2$
PROVE: Q.E.D.
⟨3⟩1. G is added to *history* at time t_1.
PROOF: Follows from step ⟨2⟩1 and the algorithm. □
⟨3⟩2. Q.E.D.
PROOF: When $fork(\phi)$ is invoked at time t_2 the conditional in line 31 will become true because of step ⟨2⟩1. Hence, $detected(\phi)$ will be invoked. □
⟨2⟩4. ASSUME: $t_1 > t_2$
PROVE: Q.E.D.
⟨3⟩1. At time t_1, $\phi \in S$.
PROOF: Follows from the assumption and steps ⟨2⟩2 and ⟨1⟩1. □
⟨3⟩2. Q.E.D.
PROOF: At time t_1 (when m_j constructs G), the algorithm must pass lines 15, 18, or 23. Each of these lines is followed by a conditional which is true because of step ⟨3⟩1. Hence, $detected(\phi)$ is invoked. □
⟨2⟩5. Q.E.D.
PROOF: Follows from the fact that steps ⟨2⟩3 and ⟨2⟩4 cover all cases. □
⟨1⟩8. Q.E.D.
PROOF: From steps ⟨1⟩6 and ⟨1⟩7 follows that the algorithm satisfies the Safety and Liveness properties of predicate detection. □

A.3 Proof of Theorem 3

ASSUME: Predicate detection is solvable.
PROVE: Σ can be implemented.

⟨1⟩1. There exists a predicate detection algorithm *PD* with operations *fork* and *detected* that solves predicate detection for a global predicate ϕ on a given computation.
PROOF: Follows from assumption. □
⟨1⟩2. Consider the algorithm in Figure 3 which uses *PD*. The *output* vector of the algorithm satisfies the accuracy property of Σ.
ASSUME: $output[i]$ was changed to $s \neq \perp$.
PROVE: p_i crashed in state s.
⟨2⟩1. The change in the output happened in line 16 of the algorithm.
PROOF: From the algorithm (there is no other line in which *output* is manipulated). □
⟨2⟩2. $detected(\phi)$ was called where ϕ is "p_i crashed in state s".
PROOF: From step ⟨2⟩1 and the algorithm (no other types of predicates are given to *PD* via *fork*). □
⟨2⟩3. ϕ holds in the computation.
PROOF: Follows from steps ⟨2⟩2 and ⟨1⟩1 and the safety property of predicate detection. □
⟨2⟩4. There exists a consistent global state in the computation of the application processes where "p_i crashed in state s" holds.
PROOF: Follows from step ⟨2⟩3 and the definition of "ϕ holds in a computation". □

⟨2⟩5. Step s is the most recent step of p_i.
PROOF: Follows from step ⟨2⟩4 and the fact that processes do not execute any more steps when they have crashed. □
⟨2⟩6. Q.E.D.
PROOF: From step ⟨2⟩4 follows that p_i has crashed and from step ⟨2⟩5 we have that the indicated state s is the most recent state of p_i. □
⟨1⟩3. The *output* vector satisfies the Completeness property of Σ.
ASSUME: p_i crashes.
PROVE: Eventually $output[i]$ will permanently change to a non-⊥ value.
⟨2⟩1. For every state s which p_i has entered in the computation, eventually $fork(\phi)$ with $\phi \equiv$"p_i crashed in state s" will be invoked.
⟨3⟩1. ASSUME: s is the initial state.
PROVE: Q.E.D.
PROOF: Follows directly from the algorithm (lines 8 and 9). □
⟨3⟩2. ASSUME: s is not the initial state.
PROVE: Q.E.D.
⟨4⟩1. For s, p_i sends a control message to all monitors in line 3.
PROOF: Follows from the algorithm and the atomicity assumption of state change and control message sending. □
⟨4⟩2. Eventually, this control message is delivered to monitor m_j.
PROOF: Follows from step ⟨3⟩1 and the reliability assumption of communication channels between an application process and the monitors and fair event scheduing. □
⟨4⟩3. Q.E.D.
PROOF: The delivery mentioned in step ⟨4⟩2 must have happened in line 12. Hence, from the algorithm a corresponding call to *fork* is made in line 13. □
⟨3⟩3. Q.E.D.
PROOF: Follows from the fact that steps ⟨3⟩1 and ⟨3⟩2 cover all cases. □
⟨2⟩2. p_i crashes in some state s.
PROOF: Follows from the assumption and the fact that the crashed state must have been reached. □
⟨2⟩3. Eventually, $detected(\phi)$ will be invoked for $\phi \equiv$"p_i crashed in state s".
PROOF: From step ⟨2⟩1 we know that $\phi \in S$ and from step ⟨2⟩2 we know that ϕ holds in the computation. From the liveness property of *PD* follows that $detected(\phi)$ must eventually be invoked. □
⟨2⟩4. Q.E.D.
PROOF: From step ⟨2⟩3 and fair event scheduling follows that the algorithm reaches line 14. From the algorithm $output[i]$ is changed to $s \neq \perp$. □
⟨1⟩4. Q.E.D.
PROOF: Follows from steps ⟨1⟩2 and ⟨1⟩3 and the definition of Σ. □