The Impossibility of Boosting Distributed Service Resilience *

Paul Attie[1,3] Rachid Guerraoui[2] Petr Kouznetsov[2] Nancy Lynch[3] Sergio Rajsbaum[4]

(1) College of Computer and Information Science, Northeastern University

(2) School of Computer and Communication Sciences, EPFL

(3) MIT Computer Science and Artificial Intelligence Laboratory

(4) Instituto de Matemáticas, Universidad Nacional Autónoma de México (UNAM)

Abstract

We prove two theorems saying that no distributed system in which processes coordinate using reliable registers and f-resilient services can solve the consensus problem in the presence of $f+1$ undetectable process stopping failures. (A service is f-resilient if it is guaranteed to operate as long as no more than f of the processes connected to it fail.)

Our first theorem assumes that the given services are atomic objects*, and allows any connection pattern between processes and services. In contrast, we show that it is possible to boost the resilience of systems solving problems easier than consensus: the k-set consensus problem is solvable for $2k-1$ failures using 1-resilient consensus services. The first theorem and its proof generalize to the larger class of* failure-oblivious *services.*

Our second theorem allows the system to contain failure-aware *services, such as failure detectors, in addition to failure-oblivious services; however, it requires that each failure-aware service be connected to all processes. Thus, $f+1$ process failures overall can disable all the failure-aware services. In contrast, it is possible to boost the resilience of a system solving consensus if arbitrary patterns of connectivity are allowed between processes and failure-aware services: consensus is solvable for any number of failures using only 1-resilient 2-process perfect failure detectors.*

1 Introduction

We consider distributed systems consisting of asynchronously operating processes that coordinate using reliable multi-writer multi-reader registers and shared services. A *service* is a distributed computing mechanism that interacts with distributed processes, accepting invocations, performing internal computation steps, and delivering responses. Examples of services include:

- Shared atomic (linearizable) objects, defined by sequential type specifications [12, 15], for example, atomic read-modify-write, queue, counter, test&set, and compare&swap objects. The consensus problem can also be defined as an atomic object.
- Concurrently-accessible data structures such as balanced trees.
- Broadcast services such as totally ordered broadcast [11].
- Failure detectors, which provide processes with hints about the failure of other processes [6].[1]

Thus, our notion of a service is quite general. We define three successively more general classes of service—atomic objects, failure-oblivious services, and general (possibly failure-aware) services—in Sections 2, 6, and 7. We define our services to tolerate a certain number f of failures: a service is f-resilient if it is guaranteed to operate as long as no more than f of the processes connected to it fail.

A fundamental, general question in distributed computing theory is: "What problems can be solved by distributed systems, with what levels of resilience, using services of given types and levels of resilience?" In this paper, we expose a basic limitation on the achievable resilience, namely, that the resilience of a system cannot be "boosted" above that of its services. More specifically, we prove two theorems saying that no distributed system in which processes coordinate using reliable registers and f-resilient services can solve the *consensus problem* in the presence of $f+1$ process stopping failures.

We focus on the consensus problem because it has been shown to be fundamental to the study of resilience in distributed systems. For example, Herlihy has shown that con-

*Paul Attie was supported by the National Science Foundation under Grant No. 0204432. Sergio Rajsbaum was supported by UNAM-PAPIIT research grant.

[1]Our notion of service encompasses all failure detectors defined by Chandra et al. [5] with one exception: we exclude failure detectors that can guess the future.

sensus is universal [12]: an atomic object of any sequential type can be implemented in a *wait-free* manner (i.e., tolerating any number of failures), using wait-free consensus objects.

Our first main theorem, Theorem 1, assumes that the given services are *atomic objects* and allows any connection pattern between processes and services. The result is a strict generalization of the classical impossibility result of Fischer et al. [9] for fault-tolerant consensus. Our simple, self-contained impossibility proof is based on a bivalence argument similar to the one in [9]. The proof involves showing that decisions can be made in a particular way, described by a *hook* pattern of executions.

In contrast to the impossibility of boosting for consensus, we show that it *is* possible to boost the resilience of systems solving problems easier than consensus. In particular, we show that the k-set consensus problem [7] is solvable for $2k-1$ failures using 1-resilient consensus services.

Theorem 1 and its proof assume that the given services are atomic objects; however, they extend to the larger class of *failure-oblivious* services. A failure-oblivious service generalizes an atomic object by allowing an invocation to trigger multiple processing steps instead of just one, and to trigger any number of responses, at any endpoints. The service may also include background processing tasks, not related to any specific endpoint. The key constraint is that no step may depend on explicit knowledge of failure events. We define the class of failure-oblivious services, give examples (e.g., totally-ordered broadcast), and describe how Theorem 1 can be extended to such services.

Our second main theorem, Theorem 11, addresses the case where the system may contain *failure-aware* services (e.g., failure detectors), in addition to failure-oblivious services and reliable registers. This result also says that boosting is impossible. However, it requires the additional assumption that each failure-aware service is connected to all processes; thus, $f+1$ process failures overall can disable all the failure-aware services. The proof is an extension of the first proof, using the same "hook" construction. We also show that the stronger connectivity assumption is necessary, by demonstrating that it *is* possible to boost the resilience of a system solving consensus if arbitrary connection patterns are allowed between processes and failure-aware services: specifically, consensus is solvable for any number of failures using only 1-resilient 2-process perfect failure detectors. The proofs of all results are available in the full version of the paper [1].

Related work. Our Theorem 1, for atomic services, can be derived by carefully combining several earlier theorems, including Herlihy's result on universality of consensus [12], and the result of Chandra et al. on f-resiliency vs. wait-freedom [4]. However, this argument does not extend to prove impossibility of boosting for failure-oblivious and failure-aware services. Moreover, some of the proofs upon which this alternative proof rests are themselves more complex than our direct proof.

Theorem 1 appeared first in a technical report [2]. Subsequent impossibility results for atomic objects appeared in [10, 16]. Our models for failure-oblivious services and general services are new. As far as we know, this is the first time a unified framework has been used to express atomic and non-atomic objects. Moreover, this is the first time boosting analysis has been performed for services more general than atomic objects.

Organization. Section 2 presents definitions for the underlying model of concurrent computation and for atomic objects. Section 3 presents our model for a system whose services are atomic objects. Section 4 presents the first impossibility result. Section 5 shows that boosting is possible for set consensus. Section 6 defines failure-oblivious services, gives an example, and extends the first impossibility result to systems with failure-oblivious services. Section 7 defines general services, gives examples, and presents our second main impossibility result. Section 8 shows how to model some important services in our framework, and Section 9 concludes.

2 Mathematical Preliminaries

2.1 Model of concurrent computation

We use the I/O automaton model [17, chapter 8] as our underlying model for concurrent computation. We assume the terminology of [17, chapter 8]. An I/O automaton A is *deterministic* iff, for each task e of A, and each state s of A, there is at most one transition (s, a, s') such that $a \in e$.

An execution α of A is *fair* iff for each task e of A: (1) if α is finite, then e is not enabled in the final state of α, and (2) if α is infinite, then α contains either infinitely many actions of e, or infinitely many occurrences of states in which e is not enabled. A *trace* of A is a sequence of external actions of A obtained by removing the states and internal actions from an execution of A. A trace of a fair execution is called a *fair trace*. If α and α' are execution fragments of A (with α finite) such that α' starts in the last state of α, then the concatenation $\alpha \cdot \alpha'$ is defined, and is called an *extension* of α.

2.2 Sequential types

We define the notion of a "sequential type", in order to describe allowable sequential behavior of atomic services. The definition used here generalizes the one in [17, chapter 9]: here, we allow nondeterminism in the choice of the initial state and the next state. Namely, sequential type $\mathcal{T} = \langle V, V_0, invs, resps, \delta \rangle$ consists of:

- V, a nonempty set of *values*,
- $V_0 \subseteq V$, a nonempty set of *initial values*,
- *invs*, a set of *invocations*,
- *resps*, a set of *responses*, and
- δ, a binary relation from $invs \times V$ to $resps \times V$ that is *total*, in the sense that, for every $(a, v) \in invs \times V$, there is at least one $(b, v') \in resps \times V$ such that $((a, v), (b, v')) \in \delta$.

We sometimes use dot notation, writing $\mathcal{T}.V, \mathcal{T}.V_0, \mathcal{T}.invs, \ldots$ for the components of $\mathcal{T}$. We say that $\mathcal{T}$ is *deterministic* if V_0 is a singleton set $\{v_0\}$, and δ is a mapping, that is, for every $(a, v) \in invs \times V$, there is *exactly one* $(b, v') \in resps \times V$ such that $((a, v), (b, v')) \in \delta$.

We allow nondeterminism in our definition of a sequential type in order to make our notion of "service" as general as possible. In particular, the problem of k-set-consensus can be specified using a nondeterministic sequential type.

Example. Read/write *sequential type:* Here, V is a set of "values", $V_0 = \{v_0\}$, where v_0 is a distinguished element of V, $invs = \{read\} \cup \{write(v) : v \in V\}$, $resps = V \cup \{ack\}$, and $\delta = \{((read, v), (v, v)) : v \in V\} \cup \{((write(v), v'), (ack, v)) : v, v' \in V\}$.

Example. Binary consensus *sequential type:* Here, $V = \{\{0\}, \{1\}, \emptyset\}$, $V_0 = \{\emptyset\}$, $invs = \{init(v)) : v \in \{0, 1\}\}$, $resps = \{decide(v) : v \in \{0, 1\}\}$, and $\delta = \{((init(v), \emptyset), (decide(v), \{v\})) : v \in V\} \cup \{((init(v), \{v'\}), (decide(v'), \{v'\})) : v, v' \in V\}$

Example. k-consensus *sequential type:* Now V is the set of subsets of $\{0, 1, \ldots, k\}$ having at most k elements, $V_0 = \{\emptyset\}$, $invs = \{init(v) : v \in \{0, 1, \ldots, k\}\}$, $resps = \{decide(v) : v \in \{0, 1, \ldots, k\}\}$, and $\delta = \{((init(v), W), (decide(v'), W \cup \{v\})) : |W| < k, v' \in W \cup \{v\}\} \cup \{((init(v), W), (decide(v'), W)) : |W| = k, v' \in W\}$.
Thus, the first k values are remembered, and every operation returns one of these values.

2.3 Canonical f-resilient atomic objects

A "canonical f-resilient atomic object" describes the allowable concurrent behavior of atomic objects. Namely, we define the *canonical f-resilient atomic object of type $\mathcal{T}$ for endpoint set J and index k*, where

- $\mathcal{T}$ is a sequential type,
- J is a finite set of *endpoints* at which invocations and responses may occur,
- $f \in \mathbf{N}$ is the level of resilience, and
- k is a unique index (name) for the service.

The object is described as an I/O automaton, in Figure 1.

The parameter J allows different objects to be connected to the same or different sets of processes. A process at endpoint $i \in J$ can issue any invocation specified by the underlying sequential type and can (potentially) receive any allowable response. We allow concurrent (overlapping) operations, at the same or different endpoints. The object preserves the order of concurrent invocations at the same endpoint i by keeping the invocations and responses in internal FIFO buffers, two per endpoint (one for invocations from the endpoint, the other for responses to the endpoint). The object chooses the result of an operation nondeterministically, from the set of results allowed by the transition relation $\mathcal{T}.\delta$ applied to the invocation and the current value of *val*. The object can exhibit nondeterminism due to nondeterminism of sequential type $\mathcal{T}$, and due to interleavings of steps for different invocations.

We model a failure at an endpoint i by an explicit input action $fail_i$. We use the task structure of I/O automata and the basic definition of fair executions to specify the required resilience: For every process $i \in J$, we assume the service has two tasks, which we call the i-perform task and i-output task. The i-perform task includes the $perform_{i,k}$ action, which carries out operations invoked at endpoint i. The i-output task includes all the $b_{i,k}$ actions giving responses at i. In addition, every i-* task ($*$ is perform or output) contains a special $dummy\_*_{i,k}$ action, which is enabled when either process i has failed or more than f processes in J have failed. The $dummy\_*_{i,k}$ action is intended to allow, but not force, the service to stop performing steps on behalf of process i after i fails or after the resilience level has been exceeded.

The definition of fairness for I/O automata says that each task must get infinitely many turns to take steps. In this context, this implies that, for every $i \in J$, the object eventually responds to an outstanding invocation at i, unless either i fails or more than f processes in J fail. If i does fail or more than f processes in J fail, the fairness definition allows the object to perform the $dummy\_*_{i,k}$ action every time the $i - *$ task gets a turn, which permits the object to avoid responding to i. In particular, if more than f processes fail, the object may avoid responding to any process in J, since $dummy\_ouput_{i,k}$ is enabled for all $i \in J$. Also, if all processes connected to the service (i.e., all processes in J) fail, the object may avoid responding to any process.

Thus, the basic fairness definition expresses the idea that the object is f-resilient: Once more than f of the processes connected to the object fail, the object itself may "fail" by becoming silent. However, although the object may stop responding, it never violates its safety guarantees, that is, it never returns values inconsistent with the underlying sequential type specification.

CanonicalAtomicObject($\mathcal{T}$, J, f, k),
where $\mathcal{T} = \langle V, V_0, invs, resps, \delta \rangle$

Signature:
Inputs:
$a_{i,k}$, $a \in invs$, $i \in J$, the invocations at endpoint i
$fail_i$, $i \in J$

Outputs:
$b_{i,k}$, $b \in resps$, $i \in J$, the responses at endpoint i

Internals:
$perform_{i,k}$, $i \in J$
$dummy\_*_{i,k}$, $* \in \{perform, output\}$, $i \in J$

State components:
val $\in V$, initially an element of V_0
inv-buffer, a mapping from J to finite sequences of *invs*, initially identically empty
resp-buffer, a mapping from J to finite sequences of *resps* initially identically empty
failed $\subseteq J$, initially $\emptyset$

Transitions:
Input: $a_{i,k}$
Effect:
add a to end of *inv-buffer*(i)

Internal: $perform_{i,k}$
Precondition:
$a = head(\textit{inv-buffer}(i))$
$\delta((a, val), (b, v))$
Effect:
remove head of *inv-buffer*(i)
val $\leftarrow v$
add b to end of *resp-buffer*(i)

Output: $b_{i,k}$
Precondition:
$b = head(\textit{resp-buffer}(i))$
Effect:
remove head of *resp-buffer*(i)

Input: $fail_i$
Effect:
failed $\leftarrow$ *failed* $\cup \{i\}$

Internal: $dummy\_*_{i,k}$
Precondition:
$i \in failed \vee |failed| > f \vee failed = J$
Effect:
none

Tasks:
For every $i \in J$:
i-perform: $\{perform_{i,k}, dummy\_perform_{i,k}\}$
i-output: $\{b_{i,k} : b \in resps\} \cup \{dummy\_output_{i,k}\}$

Figure 1. A canonical atomic object.

A canonical atomic object whose sequential type is read/write is called a *canonical register*. In this paper, we will consider canonical reliable (wait-free) registers.

2.4 f-resilient atomic objects

An I/O automaton A is an *f-resilient atomic object* of type $\mathcal{T}$ for endpoint set J and index k, provided that it implements the canonical f-resilient atomic object S of type $\mathcal{T}$ for J and k, in the following sense:

1. A and S have the same input actions (including *fail* actions) and the same output actions.
2. Any trace of A is also a trace of S. (This implies that A guarantees atomicity.)
3. Any fair trace of A is also a fair trace of S. (This says that A is f-resilient.)

We say that A is *wait-free* (or, *reliable*), if it is $(|J| - 1)$-resilient. This is equivalent to saying that (a) A is $|J|$-resilient, or (b) A is f-resilient for some $f \geq |J| - 1$, or (c) A is f-resilient for every $f \geq |J| - 1$.

3 System Model with Atomic Objects

Our system model consists of a collection of process automata, reliable registers, and fault-prone atomic objects (which we sometimes refer to as *services*). For this section, we fix I, K, and R, finite (disjoint) index sets for processes, services, and registers, respectively, and $\mathcal{T}$, a sequential type, representing the problem the system is intended to solve. A *distributed system* for I, K, R, and $\mathcal{T}$ is the composition of the following I/O automata (see [17, chapter 8]):

1. *Processes* P_i, $i \in I$,
2. *Services (atomic objects)* S_k, $k \in K$. We let $\mathcal{T}_k$ denote the sequential type, and $J_k \subseteq I$ the set of endpoints, of service S_k. We assume k itself is the index.
3. *Registers* S_r, $r \in R$. We let V_r denote the value set and $v_{0,r}$ the initial value for register S_r. We assume r is the index.

Processes interact only via services and registers. Process P_i can invoke an operation on service S_k provided that $i \in J_k$. Process P_i can also invoke a read or write operation on register S_r provided that $i \in J_r$. Services and registers do not communicate directly with one another, but may interact indirectly via processes. In the remainder of this section, we describe the components in more detail and define terminology needed for the results and proofs.

3.1 Processes

We assume that process P_i, $i \in I$ has the following inputs and outputs:

- Inputs a_i, $a \in \mathcal{T}.invs$, and outputs b_i, $b \in \mathcal{T}.resps$. These represent P_i's interactions with the external world.
- For every service S_k such that $i \in J_k$, outputs $a_{i,k}$, $a \in \mathcal{T}_k.invs$, and inputs $b_{i,k}$, $b \in \mathcal{T}_k.resps$.
- For every register S_r, outputs $a_{i,r}$, where a is a read or write invocation of S_r, and inputs $b_{i,r}$, where b is a response of S_r.

- Input $fail_i$.

P_i may issue several invocations, on the same or different services or registers, without waiting for responses to previous invocations. The external world at P_i may also issue several invocations to P_i without waiting for responses. As a technicality, we assume that when P_i performs a $decide(v)_i$ output action, it records the decision value v in a special state component.

We assume that P_i has only a single task, which therefore consists of all the locally-controlled actions of P_i. We assume that in every state, some action in that single task is enabled. We assume that the $fail_i$ input action affects P_i in such a way that, from that point onward, no output actions are enabled. However, other locally-controlled actions may be enabled—in fact, by the restriction just above, some such action *must* be enabled. This action might be a "dummy" action, as in the canonical resilient atomic objects defined in Section 2.3.

3.2 The complete system

The complete system $\mathcal{C}$ is constructed by composing the P_i, S_k, and S_r automata and then hiding all the actions used to communicate among them.

For any action a of $\mathcal{C}$, we define the *participants* of action a to be the set of automata with a in their signature. Note that no two distinct registers or services participate in the same action a, and similarly no two distinct processes participate in the same action. Furthermore, for any action a, the number of participants is at most 2. Thus, if an action a has two participants, they must be a process and either a service or register.

As we defined earlier, each process P_i has a single task, consisting of all the locally controlled actions of P_i. Each service or register S_c, $c \in K \cup R$, has two tasks for each $i \in J_c$: i-perform, consisting of $\{perform_{i,c}, dummy\_perform_{i,c}\}$, and i-output, consisting of $\{b_{i,c} : b \in \mathcal{T}_c.resps\} \cup \{dummy\_output_{i,c}\}$. These tasks define a partition of the set of all actions in the system, except for the inputs of the process automata that are not outputs of any other automata, namely, the invocations by the external world and the $fail_i$ actions. The I/O automata fairness assumptions imply that each of these tasks get infinitely many turns to execute.

We say that a task e is *applicable* to a finite execution α iff some action of e is enabled in the last state of α.

3.3 The consensus problem

The "traditional" specification of f-resilient binary consensus is given in terms of a set $\{P_i, i \in I\}$ of processes, each of which starts with some value v_i in $\{0, 1\}$. Processes are subject to stopping failures, which prevent them from producing any further output.[2] As a result of engaging in a consensus algorithm, each nonfaulty process eventually "decides" on a value from $\{0, 1\}$. The behavior of processes is required to satisfy the following conditions (see, e.g., [17, chapter 6]):

Agreement No two processes decide on different values.

Validity Any value decided on is the initial value of some process.

Termination In every fair execution in which at most f processes fail, all nonfaulty processes eventually decide.

In this paper, we specify the consensus problem differently: We say that a distributed system S *solves f-resilient consensus for I* if and only if S is an f-resilient atomic object of type consensus (Section 2.2) for endpoint set I. We argue that any system that satisfies our definition satisfies a slight variant of the traditional one. In this variant, inputs arrive explicitly via *init*() actions, not all nonfaulty processes need receive inputs, and only nonfaulty processes that do receive inputs are guaranteed to eventually decide. Our agreement and validity conditions are the same as before; our new termination condition is:

Termination In every fair execution in which at most f processes fail, any nonfaulty process *that receives an input* eventually decides.

4 Impossibility of Boosting for Atomic Objects

Our first main theorem is:

Theorem 1 *Let $n = |I|$ be the number of processes, and let f be an integer such that $0 \leq f < n-1$. There does not exist an $(f+1)$-resilient n-process implementation of consensus from canonical f-resilient atomic objects and canonical reliable registers.*

To prove Theorem 1, we assume that such an implementation exists and derive a contradiction. Let $\mathcal{C}$ denote the complete system, that is, the composition of the processes P_i, $i \in I$, services S_k, $k \in K$, and registers S_r, $r \in R$. By assumption, $\mathcal{C}$ satisfies the agreement, validity and termination properties of consensus.

For each component $c \in K \cup R$ and $i \in J_c$ (recall that J_c denotes the endpoints of c) let $inv\text{-}buffer(i)_c$ denote the invocation buffer of c, which stores invocations from P_i, and let $resp\text{-}buffer(i)_c$ denote the response buffer of c, which stores responses to P_i. Also let $buffer(i)_c = \langle inv\text{-}buffer(i)_c, resp\text{-}buffer(i)_c \rangle$.

[2] Stopping failures are usually defined as disabling the process from executing at all. However, the two definitions are equivalent with respect to overall system behavior.

4.1 Assumption

To prove Theorem 1, we make the following assumption:

(i) We assume that the processes P_i, $i \in I$, are deterministic automata, as defined in Section 2.1. For services, we assume a slightly weaker condition: that the sequential type is deterministic, i.e, the sequential type has a unique initial value and the transition relation δ is a mapping. Note that the sequential type for registers is also deterministic, by definition.

Assumption (i) implies that, after a finite *failure-free* execution α, an applicable task e determines a unique transition, arising from running task e from the final state s of α. We denote this transition as $transition(e, s)$ (since it is uniquely defined by the final state s). If $transition(e, s) = (s, a, s')$, then we write $first(e, s)$, $action(e, s)$, and $last(e, s)$ to denote s, a, and s', respectively. We sometimes abbreviate $last(e, s)$ as $e(s)$. Note that, if s is the final state of α, then $transition(e, s)$, $first(e, s)$, $action(e, s)$, and $last(e, s)$ are defined iff e is applicable to α.

Assumption (i) implies that any *failure-free* execution can be defined by applying a sequence of tasks, one after the other, to the initial state of $\mathcal{C}$. Assumption (i) does not reduce the generality of our impossibility result, because any candidate system could be restricted to satisfy (i); if the impossibility result holds for the restricted automaton, then it also holds for the original one.

Lemma 2 *Let α be any finite failure-free execution of $\mathcal{C}$, e be any task of $\mathcal{C}$ applicable to α, and $\alpha \cdot \beta$ be any failure-free extension of α such that β includes no actions of e. Then e is applicable to $\alpha \cdot \beta$.*

Let s be any state of $\mathcal{C}$ arising after a finite failure-free execution α of $\mathcal{C}$, and let e be a task that is applicable to α (equivalently, enabled in s). Then we write *participants*(e, s) for the set of participants of action $action(e, s)$. Note that, for any task e and any state s, $|participants(e, s)| \leq 2$. Also, if $|participants(e, s)| = 2$, then *participants*(e, s) is of the form $\{P_i, S_c\}$, for some $i \in I$ and $c \in K \cup R$.

4.2 Initializations and valence

In our proof, we consider executions in which consensus inputs arrive from the external world at the beginning of the execution. Thus, we define an *initialization* of $\mathcal{C}$ to be a finite execution of $\mathcal{C}$ containing exactly one *init*$()_i$ action for each $i \in I$, and no other actions. An execution α of $\mathcal{C}$ is *input-first* if it has an initialization as a prefix, and contains no other *init*$()$ actions. A finite failure-free input-first execution α is defined to be 0-*valent* if (1) some failure-free extension of α contains a *decide*$(0)_i$ action, for some $i \in I$, and (2) no failure-free extension of α contains a *decide*$(1)_i$ action, for any $i \in I$. The definition of a 1-*valent* execution is symmetric. A finite failure-free input-first execution α is *univalent* if it is either 0-valent or 1-valent. A finite failure-free input-first execution α is *bivalent* if (1) some failure-free extension of α contains a *decide*$(0)_i$ action, for some i, and (2) some failure-free extension of α contains a *decide*$(1)_i$ action, for some i. These definitions immediately imply the following result:

Lemma 3 *Every finite failure-free input-first execution of $\mathcal{C}$ is either bivalent or univalent.*

The following lemma provides the first step of the impossibility proof:

Lemma 4 *$\mathcal{C}$ has a bivalent initialization.*

For the rest of this section, fix α_b to be any particular bivalent initialization of $\mathcal{C}$.

4.3 The graph $G(\mathcal{C})$

Now define an edge-labeled directed graph $G(\mathcal{C})$ as follows:

(1) The vertices of $G(\mathcal{C})$ are the finite failure-free input-first extensions of the bivalent initialization α_b.

(2) $G(\mathcal{C})$ contains an edge labeled with task e from α to α' provided that $\alpha' = e(\alpha)$.

By assumption (i) of Section 4.1, any task triggers at most one transition after a failure-free execution α. Therefore, for any vertex α of $G(\mathcal{C})$ and any task e, there is at most one edge labeled with e outgoing from α.

4.4 The existence of a hook

We show that decisions in $\mathcal{C}$ can be made in a particular way, described by a *hook* pattern of executions. Similarly to [5], we define a *hook* to be a subgraph of $G(\mathcal{C})$ of the form depicted in Figure 2.

Lemma 5 *$G(\mathcal{C})$ contains a hook.*

4.5 Similarity

In this section, we introduce notions of similarity between system states. These will be used in showing nonexistence of a hook, which will yield the contradiction needed for the impossibility proof.

Let $j \in I$ and let s_0 and s_1 be states of $\mathcal{C}$. Then s_0 and s_1 are j-*similar* if:

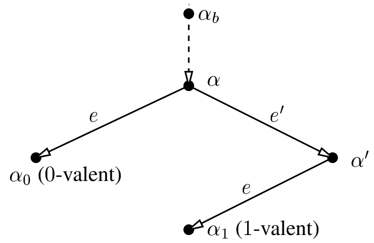

Figure 2. A hook starting in α.

(1) For every $i \in I - \{j\}$, the state of P_i is the same in s_0 and s_1.

(2) For every $c \in K \cup R$:

1. The value of val_c is the same in s_0 and s_1.
2. For every $i \in J_c - \{j\}$, the value of $buffer(i)_c$ is the same in s_0 and s_1.

Lemma 6 *Let $j \in I$. Let α_0 and α_1 be finite failure-free input-first executions, s_0 and s_1 the respective final states of α_0 and α_1. Suppose that s_0 and s_1 are j-similar. If α_0 and α_1 are univalent, then they have the same valence.*

Similarly, we define the notion of k-similar states: Let $k \in K$, and let s_0 and s_1 be states of $\mathcal{C}$. Then s_0 and s_1 are *k-similar* if:

(1) For every $i \in I$, the state of P_i is the same in s_0 and s_1.

(2) For every $c \in (K - \{k\}) \cup R$, the state of S_c is the same in s_0 and s_1.

Lemma 7 *Let $k \in K$. Let α_0 and α_1 be finite failure-free input-first executions, s_0 and s_1 the respective final states of α_0 and α_1. Suppose that s_0 and s_1 are k-similar. If α_0 and α_1 are univalent, then they have the same valence.*

4.6 The non-existence of a hook

Now we are ready to prove the absence of hooks. We assume that a hook exists, and we locate in $G(\mathcal{C})$ two univalent executions of opposite valence that produce j-similar ($j \in J$) or k-similar states ($k \in K$). Lemmas 6 and 7 establish a contradiction.

Lemma 8 $G(\mathcal{C})$ *contains no hooks.*

Lemma 5 contradicts Lemma 8. Hence we have derived a contradiction by assuming the negation of Theorem 1. Hence Theorem 1 is established.

5 k-Set Consensus

Our boosting impossibility result concerns *consensus* implementations. Interestingly, while it is not possible to implement $(f+1)$-resilient *consensus* using registers and f-resilient atomic objects, this is not the case for the *k-set consensus problem* [7]. In k-set consensus, the processes have to agree on at most k different values (k-set consensus reduces to consensus when $k = 1$).

Consider a set of f-resilient k-set consensus services, each one exporting *m ports*. An algorithm that implements f'-resilient k'-set consensus works as follows. Take a *principal* subset of the processes, and divide it into s disjoint groups, each one accessing a different service. Each principal process participates in an execution proposing its input value to its designated service. When it gets a decision back, the process decides on the value and writes it in a shared register. The remaining processes simply wait until at least one principal process writes the value. The values of k' and f' depend on the size of the principal set, and on the number s of services we divide it into. There is a tradeoff between k' and f': if a small number of failures f' is tolerated, then a high degree of agreement is achieved, namely a small k'. If more failures f' must be tolerated, then a lower degree of agreement is achieved, namely a large k'.

To achieve correctness, we must ensure first that at least one principal process receives a decision from its service and communicates the decision to all, i.e., (1) every f-resilient service is connected to $f+1$ processes, and (2) fewer than $s \cdot (f+1)$ principal processes can fail: $f' < s \cdot (f+1)$. Thus, there is at least one service S that is not killed, and moreover, there is at least one correct principal process that receives a decision value from S and writes the decision in a shared register. Thus, every correct process eventually decides. The number of possible different decision values is at most $s \cdot k$: there are at most k different values returned per service; more precisely, at most k values per service being accessed by at least k processes, and c values for a service that is being accessed by c processes for $c < k$. Thus, for a desired overall resilience f', we want the smallest possible k' and so we find the smallest integer s that guarantees $f' < s \cdot (f+1)$. Thus, we have $s = \lceil (f'+1)/(f+1) \rceil$ services, and take the first $f'+1$ processes to be the principal processes ($f'+1$ processes using as few services as possible, each one with $f+1$ input ports). It follows that

Theorem 9 *For any $1 \leq k < m$, $k \leq f \leq m-1$, $1 \leq f' \leq n-1$, it is possible to implement f'-resilient k'-set consensus using read-write memory and f-resilient k-set consensus services, each one with m ports, for*

$$k' \geq k \cdot \left\lfloor \frac{f'+1}{f+1} \right\rfloor + \min(k,\ (f'+1) \bmod (f+1)).$$

When each available service is wait-free, that is $f = m - 1$, this algorithm reduces to the one of [13], and gives a tight bound. As an example, assume that we want to implement a f'-resilient k'-set consensus in a system of $2c$ processes, where $f' = 2c - 1$, using only 1-resilient consensus services, i.e., $f = 1$, $k = 1$. The smallest k' for which we can do this is $k' = c$, using $s = c$ services, each shared by 2 processes ($f' + 1 = 2c$ principal processes).

Note that the algorithm above uses services that are not connected to all processes. It is known that f-resilient f-set consensus cannot be solved using only reliable registers [3, 14, 18]. We conjecture that f-resilient f-set consensus cannot be solved using only reliable registers and services that are connected to all processes.

6 Failure-Oblivious Services

A *failure-oblivious service* is a generalization of an atomic object. It allows an invocation to trigger multiple processing steps instead of just one *perform* step. These steps can interleave with processing steps triggered by other invocations, and this makes a failure-oblivious service non-atomic, in general. A failure-oblivious service also allows an invocation to trigger any number of responses, at any endpoints, instead of just a single response at the endpoint of the invocation. The service may also include background processing tasks, not related to any specific endpoint. The key constraint is that no step may depend on explicit knowledge of failure events. In this section, we define the class of failure-oblivious services, give examples, and describe how Theorem 1 can be extended to such services.

6.1 f-resilient failure-oblivious services

As for atomic objects, we begin by defining a canonical f-resilient failure-oblivious service. A *canonical f-resilient failure-oblivious service* is parameterized by J, f, and k, which have the same meanings as for canonical atomic objects. Also, in place of the sequential type parameter $\mathcal{T}$, the service has a *service type parameter $\mathcal{U}$*, which is a tuple $\langle V, V_0, \mathit{invs}, \mathit{resps}, \mathit{glob}, \delta_1, \delta_2, \delta_3 \rangle$, where V and V_0 are as before, *invs* and *resps* are the respective sets of invocations and responses (which can occur at any endpoint), *glob* is a set of *global tasks*, and $\delta_1, \delta_2, \delta_3$ are three transition relations.

Here, δ_1 is a total binary relation from $\mathit{invs} \times J \times V$ to (the set of mappings from J to finite sequences of *resps*) $\times V$. It is used to map an invocation at the head of a particular *inv-buffer*, and the current value for val, to a set of results, each of which consists of a new value for val and sequences of responses to be added to any or all of the *resp-buffer*s. δ_2 is a total binary relation from $J \times V$ to (the set of mappings from J to finite sequences of *resps*) $\times V$. It is used to map a particular endpoint and value of val to a set of results, defined as above. Finally, δ_3 is a total binary relation from V to (the set of mappings from J to finite sequences of *resps*) $\times V$. It it used to map a value of val to a set of results. The code for a canonical failure-oblivious automaton, showing how these parameters are used, appears in Figure 3.

Thus, a canonical f-resilient failure-oblivious service is allowed to perform rather flexible kinds of processing, both related and unrelated to individual endpoints, as long as processing decisions do not depend on knowledge of occurrence of failure events.

An I/O automaton A is an *f-resilient failure-oblivious service* of type $\mathcal{U}$, endpoint set J, and index k, provided that it implements the canonical f-resilient failure oblivious service S of type $\mathcal{U}$ for J and k, in the same sense as for atomic objects.

6.2 Impossibility of Boosting

Let index set K include now the indices of all failure-oblivious services. Now the notion of k-similarity restricts the states of all registers and of all atomic and failure-oblivious services except S_k. We show, in the full version of the paper, that Lemmas 2–8 extend to this case. Hence the following result:

Theorem 10 *Let f and n be integers, $0 \leq f < n-1$. There does not exist an $(f+1)$-resilient n-process implementation of consensus from canonical f-resilient atomic services, canonical f-resilient failure-oblivious services, and canonical reliable registers.*

7 General (Failure-Aware) Services

A *general*, or *failure-aware* service is a further generalization of a failure-oblivious service. This time, the generalization removes the failure-oblivious constraint, allowing the service's decisions to depend on knowledge of failures of processes connected to the service.

7.1 f-resilient general services

A *canonical f-resilient general service* is parameterized by J, f, and k, which have the same meanings as for canonical failure-oblivious services, and by a service type parameter $\mathcal{U}$, which is a tuple $\langle V, V_0, invs, resps, glob, \delta_1, \delta_2, \delta_3 \rangle$, as for failure-oblivious services. This time, however, the domains of δ_1, δ_2, and δ_3 are $\mathit{invs} \times J \times V \times 2^I$, $J \times V \times 2^I$, and $V \times 2^I$, respectively. The final argument, in each case, will be instantiated in the service code with the current *failed* set.

The only portions of the code that are different from those for failure-oblivious services are the three transition definitions that use the δ_1, δ_2, and δ_3 (Figure 4).

CanonicalFailureObliviousService$(\mathcal{U}, J, f, k)$,
where $\mathcal{U} = \langle V, V_0, \mathit{invs}, \mathit{resps}, glob, \delta_1, \delta_2, \delta_3 \rangle$

Signature:
Inputs:
$a_{i,k}$, $a \in \mathit{invs}$, $i \in J$
$fail_i$, $i \in J$

Outputs:
$b_{i,k}$, $b \in \mathit{resps}$, $i \in J$

Internals:
$perform_{i,k}$, $i \in J$
$compute_{i,k}$, $i \in J$
$dummy\_*_{i,k}$, $* \in \{perform, compute, output\}$, $i \in J$
$compute_{g,k}$, $g \in glob$
$dummy\_compute_{g,k}$, $g \in glob$

State components:
As for canonical atomic object.

Transitions:
Input: $a_{i,k}$
As for canonical atomic object.

Internal: $perform_{i,k}$
Precondition:
$a = head(\mathit{inv\text{-}buffer}(i))$
$\delta_1((a, i, \mathit{val}), (B, v))$
Effect:
remove head of *inv-buffer*(i)
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Internal: $compute_{i,k}$, $i \in J$
Precondition:
$\delta_2((i, \mathit{val}), (B, v))$
Effect:
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Internal: $compute_{g,k}$, $g \in glob$
Precondition:
$\delta_3(\mathit{val}, (B, v))$
Effect:
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Output: $b_{i,k}$
As for canonical atomic object.

Input: $fail_i$
As for canonical atomic object.

Internal: $dummy\_*_{i,k}$, $i \in J$
As for canonical atomic object.

Internal: $dummy\_compute_{g,k}$, $g \in glob$
Precondition:
$|\mathit{failed}| > f$
Effect:
none

Tasks:
For every $i \in J$:
$i\text{-}perform$: $\{perform_{i,k}, dummy\_perform_{i,k}\}$
$i\text{-}compute$: $\{compute_{i,k}, dummy\_compute_{i,k}\}$
$i\text{-}output$: $\{b_{i,k} : b \in \mathit{resps}\} \cup \{dummy\_output_{i,k}\}$
For every $g \in glob$:
$g\text{-}compute$: $\{compute_{g,k}, dummy\_compute_{g,k}\}$

Figure 3. A canonical failure-oblivious service.

Internal: $perform_{i,k}$
Precondition:
$a = head(\mathit{inv\text{-}buffer}(i))$
$\delta_1((a, i, \mathit{val}, \mathit{failed}), (B, v))$
Effect:
remove head of *inv-buffer*(i)
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Internal: $compute_{i,k}$, $i \in J$
Precondition:
$\delta_2((i, \mathit{val}, \mathit{failed}), (B, v))$
Effect:
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Internal: $compute_{g,k}$, $g \in glob$
Precondition:
$\delta_3((\mathit{val}, \mathit{failed}), (B, v))$
Effect:
$\mathit{val} \leftarrow v$
for $j \in J$ do
add $B(j)$ to end of *resp-buffer*(j)

Figure 4. Relations δ_1, δ_2 and δ_3 in a general service.

An I/O automaton A is an f*-resilient general service* of type $\mathcal{U}$, endpoint set J, and index k, provided that it implements the canonical f-resilient general service S of type $\mathcal{U}$ for J and k, in the same sense as for atomic and failure-oblivious services.

7.2 Impossibility of Boosting

Our impossibility results for atomic and failure-oblivious services allow arbitrary connections between processes and services. However, it turns out that we *can* boost the resilience of systems containing failure-aware services, if we allow arbitrary connection patterns:

For example, consider a system that uses wait-free registers and 1-resilient perfect failure detectors. Suppose that every pair of processes shares a 1-resilient 2-process failure detector. Such a system can implement a *wait-free* perfect failure detector for all processes as follows: Process i just listens to all failure detectors it is connected to and accumulates the set of suspected processes in a dedicated register. Periodically, it outputs its set of suspected processes. Since every perfect failure detector is 1-resilient, the algorithm is wait-free. Using this construction, f-resilient consensus, for any f, can be implemented using wait-free registers and 1-resilient services.

This boosting is, however, impossible if we assume a system in which f-resilient failure-aware services must be connected to all processes, thus, $f + 1$ process failures overall can disable all the failure-aware services. We assume that the system may also contain f-resilient failure-oblivious services, connected to arbitrary processes. By applying arguments similar to ones presented in Section 4, we can prove boosting to be impossible, i.e., that $(f+1)$-resilient consensus cannot be solved in such a model.

Theorem 11 *Let f and n be integers, $0 \leq f < n-1$. There does not exist an $(f+1)$-resilient n-process implementation of consensus from canonical f-resilient general services connected to all processes, canonical f-resilient atomic services (connected to arbitrary processes), canonical f-resilient failure-oblivious services (connected to arbitrary processes), and canonical reliable registers.*

8 Examples

In the full version [1] we show how totally ordered broadcast, and various failure detectors, can be modeled in our framework. An f-resilient totally ordered broadcast service can be modeled as an f-resilient failure-oblivious service. An invocation inserts a message into a queue (which is the value *val*), and a $g\text{-}compute$ action subsequently inserts this message into every output buffer of the service. We use general (failure-aware) services to model failure detectors. Our failure detectors do not provide all the functionality of the standard model [5]: because our failure detectors are automata, they cannot predict future input actions. Thus, our services encompass only *realistic* failure detectors [8]. Otherwise, modeling failure detectors using general services is straightforward, since general services have access to the set of failed processes.

9 Conclusions

We have established the impossibility of boosting the resilience of services in a distributed asynchronous system where processes are subject to undetectable stopping failures. Our results can be viewed as a generalization to any number f of failures of the impossibility result of Fischer, Lynch and Paterson [9] for $f = 1$. While our first result (for atomic objects) can be derived from existing results in the literature, the direct proof that we give is simpler, and is also easily extended to more general services than atomic objects.

References

[1] P. C. Attie, R. Guerraoui, P. Kouznetsov, N. A. Lynch, and S. Rajsbaum. The impossibility of boosting distributed service resilience. Technical report, MIT Computer Science and Artificial Intelligence Laboratory, 2005. Available at `http://theory.lcs.mit.edu/tds/papers/Attie/boosting-tr.ps`.

[2] P. C. Attie, N. A. Lynch, and S. Rajsbaum. Boosting fault-tolerance in asynchronous message passing systems is impossible. Technical report, MIT Laboratory for Computer Science, MIT-LCS-TR-877, 2002.

[3] E. Borowsky and E. Gafni. Generalized FLP impossibility result for t-resilient asynchronous computations. In *Proceedings of the 25th ACM Symposium on Theory of Computing (STOC)*, pages 91–100, May 1993.

[4] T. Chandra, V. Hadzilacos, P. Jayanti, and S. Toueg. Wait-freedom vs. t-resiliency and the robustness of wait-free hierarchies. In *Proceedings of the 13th Annual ACM Symposium on Principles of Distributed Computing (PODC'94)*, pages 334–343, August 1994.

[5] T. D. Chandra, V. Hadzilacos, and S. Toueg. The weakest failure detector for solving consensus. *Journal of the ACM*, 43(4):685–722, July 1996.

[6] T. D. Chandra and S. Toueg. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM*, 43(2):225–267, March 1996.

[7] S. Chaudhuri. Agreement is harder than consensus: set consensus in totally asynchronous systems. In *Proceedings of the 19th Annual ACM Symposium on Principles of Distributed Computing (PODC'00)*, pages 311–324, August 1990.

[8] C. Delporte-Gallet, H. Fauconnier, and R. Guerraoui. A realistic look at failure detectors. In *IEEE Symposium on Dependable Systems and Networks (DSN 2002)*, Washington DC, June 2002.

[9] M. J. Fischer, N. A. Lynch, and M. S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(3):374–382, April 1985.

[10] R. Guerraoui and P. Kouznetsov. On failure detectors and type boosters. In *Proceedings of the 17th International Symposium on Distributed Computing (DISC'03)*, October 2003.

[11] V. Hadzilacos and S. Toueg. A modular approach to fault-tolerant broadcast and related problems. Technical report, Cornell University, Computer Science, May 1994.

[12] M. Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124–149, January 1991.

[13] M. Herlihy and S. Rajsbaum. Algebraic spans. *Mathematical Structures in Computer Science (Special Issue: Geometry and Concurrency)*, 10(4):549–573, August 2000.

[14] M. Herlihy and N. Shavit. The topological structure of asynchronous computability. *Journal of the ACM*, 46(6):858–923, November 1999.

[15] M. Herlihy and J. M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, June 1990.

[16] P. Jayanti. Private communication. 2003.

[17] N. A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers, 1996.

[18] M. Saks and F. Zaharoglou. Wait-free k-set agreement is impossible: The topology of public knowledge. *SIAM Journal on Computing*, 29:1449–1483, 2000.