A Closer Look at Fault Tolerance

Gadi Taubenfeld
The Interdisciplinary Center, P.O.Box 167, Herzliya 46150, Israel
tgadi@idc.ac.il

ABSTRACT

The traditional notion of fault tolerance requires that *all* the correct participating processes eventually terminate, and thus, is not sensitive to the *number* of correct processes that should properly terminate as a result of failures. Intuitively, an algorithm that in the presence of any number of faults always guarantees that all the correct processes except maybe one properly terminate, is more resilient to faults than an algorithm that in the presence of a single fault does not even guarantee that a single correct process ever terminates. However, according to the standard notion of fault tolerance both algorithms are classified as algorithms that can not tolerate a single fault.

To overcome this difficulty, we generalize the traditional notion of fault tolerance in a way which enables to capture more sensitive information about the resiliency of an algorithm. Then, we present several algorithms for solving classical problems which are resilient under the new notion. It is well known that, in an asynchronous systems where processes communicate either by reading and writing atomic registers or by sending and receiving messages, important problems such as, consensus, set-consensus, election, perfect renaming, implementations of a test-and-set bit, a shared stack, a swap object and a fetch-and-add object have no deterministic solutions which can tolerate even a single fault. We show that while, some of these problems have solutions which guarantee that in the presence of *any* number of faults most of the correct processes will properly terminate; other problems do not even have solutions which guarantee that in the presence of just *one* fault at least one correct process properly terminates.

Categories and Subject Descriptors

F.0 [**Theory of Computation**]: General

General Terms

Algorithms, Reliability, Theory

PODC'12, July 16–18, 2012, Madeira, Portugal.

Keywords

Fault tolerance, shared memory, message passing, election, test-and-set, renaming, consensus, set-consensus, stack, swap, fetch-and-add.

1. INTRODUCTION

1.1 Motivation

According to the standard notion of fault tolerance, an algorithm is t-resilient if in the presence of up to t faults, *all* the correct processes can still properly complete their operations and terminate. Thus, an algorithm is *not* t-resilient, if as a result of t faults there is *some* correct process that can not properly terminate. This traditional notion of fault tolerance is not sensitive to the *number* of correct processes that may or may not complete their operations as a result of the failure of other processes.

Consider for example the renaming problem, which allows processes, with distinct initial names from a large name space, to get distinct new names from a small name space. A renaming algorithm that in the presence of any number of faults always guarantees that *most* of the correct processes, but not necessarily all, get distinct new names is clearly more resilient than a renaming algorithm that in the presence of a single fault does not guarantee that even one correct process ever gets a new name. However, using the standard notion of fault tolerance, it is not possible to compare the resiliency of such algorithms – as both are simply not even 1-resilient. This motivates us to suggest and investigate a more general notion of fault tolerance.

We have generalized the traditional notion of fault tolerance by allowing a limited number participating correct processes not to terminate in the presence of faults. Every process that do terminate is required to return a correct result. Thus, our definition guarantees safety but may sacrifice liveness (termination), for a limited number of processes, in the presence of faults. The consequences of violating liveness are often less sever than those of violating safety. In fact, there are systems that can detect and abort processes that run for too long. Sacrificing liveness for few of the processes allows us to increase the resiliency of the whole system.

1.2 Model and Basic Definitions

Our model of computation consists of an asynchronous collection of n processes that communicate either by reading and writing atomic registers or by sending and receiving messages. The processes have unique identifiers. With an atomic register, it is assumed that operations on the register

occur in some definite order. That is, reading or writing an atomic register is an indivisible action.

With required participation every process must eventually execute its code. However, a more interesting and practical situation is one in which participation is not required, as is usually assumed when solving resource allocation problems. Unless explicitly stated, when the shared memory model is considered, it is assumed that participate is *not* required. When the message passing model is considered, it is assumed that: participate is required, a process starts participating spontaneously or when receiving a first message. Once a process starts participating it may fail by *crashing*.

In the literature, it is common to assume that the identifiers of the n processes are integers taken from the range $\{1, ..., n\}$. However, there may be situations when there are many more identifiers than processes. For example, there might be a small number of processes, say 50, but their identifiers can be taken from the range $\{0, ..., 2^{32}\}$. In such a case identifiers cannot be easily used to index registers, and hence it is better to use symmetric algorithms.

> **Symmetric Algorithms:** A symmetric algorithm is an algorithm in which the only way for distinguishing processes is by comparing identifiers, which are unique. Identifiers can be written, read and compared, but there is no way of looking inside any identifier. Thus, identifiers cannot be used to index shared registers.

Designing symmetric algorithms is especially important, when the designed algorithms are intended to be used as a building blocks in an environment where the processes' name space is not known in advance. Most of the algorithms presented in this paper are symmetric.

1.3 Fault Tolerance

For the rest of the paper, n denotes the number of processes, t denotes the number of faulty processes, and $N = \{0, 1, ..., n\}$.

> **Definition:** For a given function $f : N \to N$, an algorithm is (t, f)-*resilient* if in the presence of t' faults at most $f(t')$ participating correct processes may *not* properly terminate their operations, for every $0 \leq t' \leq t$.

It seems that (t, f)-*resiliency* is interesting only when requiring that $f(0) = 0$. That is, in the absence of faults all the participating processes must properly terminate. The standard definition of t-resiliency is equivalent to (t, f)-resiliency where $f(t') = 0$ for every $0 \leq t' \leq t$. Thus, the familiar notion of *wait-freedom* is equivalent to $(n-1, f)$-resiliency where $f(t') = 0$ for every $0 \leq t' \leq n-1$. The new notion of (t, f)-resiliency is quite general, and in this paper we focus mainly on the following three levels of resiliency.

- An algorithm is *almost-t-resilient* if it is (t, f)-*resilient*, for a function f where $f(0) = 0$ and $f(t') = 1$, for every $1 \leq t' \leq t$. Thus, in the presence of any number of up to t faults, all the correct participating processes, except maybe one process must properly terminate.
- An algorithm is *partially-t-resilient* if it is (t, f)-*resilient*, for a function f where $f(0) = 0$ and $f(t') = t'$, for every $1 \leq t' \leq t$. Thus, in the presence of any number $t' \leq t$ faults, all the correct participating processes, except maybe t' of them must properly terminate.
- An algorithm is *weakly-t-resilient* if it is (t, f)-*resilient*, for a function f where $f(0) = 0$, and in the presence of any number of up to $t \geq 1$ faults, if there are *two* or more correct participating processes then one correct participating process must properly terminate. (Notice that for $n = 2$, if one process fails the other one is not required to terminate.)

For $n \geq 3$ and $t < n/2$, the notion of weakly-t-resiliency is strictly weaker than the notion of partially-t-resiliency. For $n \geq 3$, the notions of weakly-t-resiliency is strictly weaker than the notion of almost-t-resiliency. For $n \geq 3$ and $t \geq 2$, the notions of partially-t-resiliency is strictly weaker than almost-t-resiliency. For all n, partially-1-resiliency and almost-1-resiliency are equivalent. For $n = 2$ these three notions are equivalent. We say that an algorithm is *almost-wait-free* if it is *almost-$(n-1)$-resilient*, thus, in the presence of any number of faults, all the participating correct processes, except maybe one process must terminate. We say that an algorithm is *partially-wait-free* if it is *partially-$(n-1)$-resilient*, thus, in the presence of any number of $t \leq n-1$ faults, all the correct participating processes, except maybe t of them must properly terminate.

In an asynchronous shared memory system which supports atomic registers or in a message passing system, important problems such as consensus, set-consensus, election, perfect renaming, implementations of a test-and-set bit, a shared stack, a swap object and a fetch-and-add object, have no solutions which can tolerate even a single fault. Rather surprisingly, as we will show later, while some of these problems have solutions which satisfy almost-wait-freedom, other problems do not even have weakly-1-resilient solutions.

1.4 Contributions

New Definitions. We generalize the traditional notion of fault tolerance. Together with the technical results, the new definitions provide a deeper understanding of complexity and computability issues which are involved in the development of fault-tolerant algorithms.

Election. In this problem one or more processes independently initiate their participation in an election to decide on a leader. Each participating process should eventually output either 0 or 1 and terminate. At most one process may output 1, and in the absence of faults exactly one of the *participating* processes should output 1. The process which outputs 1 is the elected leader. It is known that there is no 1-resilient election algorithm, when processes communicate either by reading and writing atomic registers or by sending and receiving messages. We show that:

> (1) *There is an almost-wait-free symmetric election algorithm using* $\lceil \log n \rceil + 2$ *atomic registers.*
> (2) *There is an almost-wait-free symmetric election algorithm with* $n^2 - n$ *message complexity.*

Message complexity is the total number of message sent. The known space lower bound for election in the absence of faults is $\lceil \log n \rceil + 1$ atomic registers [33].

Test-and-set. A test-and-set bit is an object that supports two operations called *test-and-set* and *reset*. A test-and-set

operation on a single bit takes as argument a shared bit b, assigns the value 1 to b, and returns the previous value of b (which can be either 0 or 1). A reset operation takes as argument a shared registers b and writes the value 0 into b. We show that:

> (1) *There is an almost-wait-free symmetric implementation of a test-and-set bit for n processes using $n+1$ atomic registers.* (2) *Any implementation of a test-and-set bit for n processes using registers must use at least n registers, even in the absence of faults.*

It is known that in asynchronous systems where processes communicate using atomic registers there are no 1-resilient implementations of a test-and-set bit [28].

Perfect Renaming. A *perfect* renaming algorithm allows n processes with initially distinct names from a large name space to acquire distinct new names from the set $\{1, ...n\}$. A *one-shot* renaming algorithm allows each process to acquire a distinct new name just once. A *long-lived* renaming algorithm allows processes to repeatedly acquire distinct names and release them. We show that:

> (1) *There is a partially-wait-free symmetric one-shot perfect renaming algorithm using* (*a*) $n-1$ *almost-wait-free election objects, or* (*b*) $O(n \log n)$ *registers, or* (*c*) $O(n^3)$ *messages.* (2) *There is a partially-wait-free symmetric long-lived perfect renaming algorithm using either* $n-1$ *almost-wait-free test-and-set bits* or $O(n^2)$ *registers.*

It is known that in asynchronous systems where processes communicate either by atomic registers or by sending and receiving messages, there is no 1-resilient perfect renaming algorithm [7, 30, 35].

Fetch-and-add, swap, stack. A *fetch-and-add* object supports an operation which takes as arguments a shared register r, and a value *val*. The value of r is incremented by *val*, and the old value of r is returned. A *swap* object supports an operation which takes as arguments a shared registers and a local register and atomically exchange their values. A shared stack is a linearizable object that supports push and pop operations, by several processes, with the usual stack semantics. We show that:

> *There are partially-wait-free implementations of a fetch-and-add object, a swap object, and a stack object using atomic registers.*

The result complements the results that in asynchronous systems where processes communicate using registers there are no 1-resilient implementations of fetch-and-add, swap, and stack objects [13, 22].

Consensus and Set-consensus. The *k-set consensus* problem is to find a solution for n processes, where each process starts with an input value from some domain, and must choose some participating process' input as its output. All n processes together may choose no more than k distinct output values. The 1-set consensus problem, is the familiar consensus problem. We show that:

> (1) *For $n \geq 3$ and $1 \leq k \leq n-2$, there is no weakly-k-resilient k-set-consensus algorithm using either atomic registers or sending and receiving messages. In particular, for $n \geq 3$, there is no weakly-1-resilient consensus algorithm using either atomic registers or messages.* (2) *For $n \geq 3$ and $1 \leq k \leq n-2$, there is no weakly-k-resilient k-set-consensus algorithm using almost-wait-free test-and-set bits and atomic registers.*

Our results strengthen the know results that, in asynchronous systems where processes communicate either by atomic registers or by sending and receiving messages, there is no 1-resilient consensus algorithm [20, 28], and there is no k-resilient k-set-consensus algorithm [12, 23, 32].

1.5 Related Work

In [33] it is proved that, in the absence of failures, $\lceil \log n \rceil + 1$ registers are necessary and sufficient for symmetric election, assuming that only the elected leader is required to ever terminate, while n registers are necessary and sufficient for deadlock-free symmetric mutual exclusion. We use some key ideas from [33], in our implementations of an almost-wait-free election object and an almost-wait-free test-and-set bit. The impossibility result that there are no election algorithm and no perfect renaming algorithm that can tolerate a single crash failure was first proved for the asynchronous message-passing model in [8, 30], and later has been extended for the shared memory model in [35].

The one-shot renaming problem was first solved for message-passing systems [8], and later for shared memory systems [11]. In [17] a long-lived wait-free renaming algorithm was presented. Several of the many papers on renaming are [1, 3, 4, 5, 9, 10, 14, 18, 21, 23, 26, 29].

The consensus problem was formally defined in [31]. The impossibility result that there is no consensus algorithm that can tolerate even a single crash failure was first proved for the asynchronous message-passing model in [20], and later has been extended for the shared memory model with atomic registers, in [28]. The impossibility result that, for $1 \leq k \leq n-1$ there is no k-resilient k-set-consensus algorithm for n processes using atomic registers, is from [12, 23, 32].

Extensions of the notion of fault tolerance, which are different from those considered in this paper, were proposed recently in[19], where a precise way is presented to characterize adversaries by introducing the notion of disagreement power: the biggest integer k for which the adversary can prevent processes from agreeing on k values when using registers only; and it is shown how to compute the disagreement power of an adversary. The ability to solve consensus under various symmetric and asymmetric progress conditions was studied in [25, 34].

A comprehensive discussion of wait-free synchronization is given in [22]. In [6], a class of objects called Common2 is defined. Each object in Common2 has a wait-free implementation from registers together with any other object in Common2. Commonly used objects such as test-and-set, fetch-and-add, swap, and stack are in Common2 [2, 6]. In [24], the related notion of a non-blocking is introduced. It guarantees that some correct process with a pending operation, will always be able to complete its operation in a finite number of its own steps regardless of the execution speed of other processes. For one-shot objects wait-freedom and non-blocking are the same.

2. ALMOST-WAIT-FREE SYMMETRIC ELECTION

In the leader election problem, processes do not have inputs. Each participating process should eventually output either 0 or 1 and terminate. At most one process may output the value 1, and in the absence of faults exactly one of the participating processes should output 1. The process which outputs 1 is elected as a leader. It is not require that the processes know the identity of the leader. The elected leader must be one of the participating processes, thus, there can not be an a priori leader.

In asynchronous systems where processes communicate either using atomic registers or by sending and receiving messages, election is impossible with one faulty process [20, 30, 35]. We show below that almost-wait-free symmetric election is possible in such asynchronous systems. This possibility result for election is later used for solving perfect renaming. We point out that, it follows from the results presented in Section 6 for the consensus problem, that for a stronger definition of election in which it is required that the processes know (i.e., output) the identity of the leader, even weakly-1-resilient strong-election is impossible.

2.1 Election using atomic registers

In [33], an election algorithm which is *not* weakly-1-resilient is presented. It is correct under the following assumptions: (1) processes never fail, and (2) only the elected leader is required to terminate. The election algorithm presented below, is based on the algorithm from [33].

THEOREM 2.1. *There is an almost-wait-free symmetric election algorithm using* $\lceil \log n \rceil + 2$ *atomic registers.*

The algorithm below is for n processes each with a unique identifier taken from some (possibly infinite) set which does not include 0. The algorithm uses the shared registers *turn* and *done* and the array of registers V. All these registers are initially 0. Also, for each process, the local variables *level* and j are used. We denote by *e.turn*, *e.done* and $e.V[*]$ the shared registers of the specific election algorithm (object) named e. This should simplify the construction of algorithms that use election as a basic building block.

AN ALMOST-WAIT-FREE SYMMETRIC ELECTION: `process p's program.`

```
function election (e: object_name) return:value in {0,1};
1   e.turn := p;
2   for level := 1 to ⌈log n⌉ do
3       repeat
4           if e.done = 1 then return(0) fi;  /*not leader*/
5           if e.turn ≠ p then
6               for j := 1 to level − 1 do
7                   if e.V[j] = p then e.V[j] := 0 fi od;
8               return(0) fi                  /* not the leader */
9       until e.V[level] = 0;
10      e.V[level] := p;
11      if e.turn ≠ p then
12          for j := 1 to level do
13              if e.V[j] = p then e.V[j] := 0 fi od;
14          return(0) fi                      /* not the leader */
15  od;
16  e.done := 1; return(1).                   /* leader! */
end_function
```

The process that is last to write to *e.turn* (line 1) attempts to become the leader. It does so, by waiting for each of the registers $e.V[j]$ to be 0 (lines 3-9) and then sets the register to its id (line 10). A process becomes the leader if it manages to write its id into all the registers during the period that *e.turn* equals its id. Any process that notices that *e.turn* is no longer equals its id, gives up on becoming the leader, and erase any write it has made (lines 6 & 12).

There are runs of the algorithm in which every process manages to set $\lceil \log n \rceil$ registers before discovering that another process has modified *e.turn*, and as a result has to set back to 0 some of the registers before terminating. Proving the correctness of the algorithm is rather challenging, due to the existence of such runs.

In [33], it has been proven that, even in the absence of faults, any election algorithm for n processes must use at least $\lceil \log n \rceil + 1$ registers. (This lower bound holds even for non-symmetric algorithms.) Thus, our algorithm which uses $\lceil \log n \rceil + 2$ registers, provides an almost tight space upper bound.

2.2 Correctness proof

The proof of the election algorithm is an adaptation of the proof for the algorithm from [33] which guarantees that only the leader terminates, and is correct only in the absence of faults. The fact that our election algorithm uses $\lceil \log n \rceil + 2$ atomic registers is obvious from inspecting the algorithm.

THEOREM 2.2 (LIVENESS). *In the absence of faults, at least one leader is elected.*

PROOF. Assume to the contrary that no leader is elected. Let r be an infinite run with no faults where no leader is elected, and let p be the last processes to write to *turn* in run r. Let q be the process with the highest value of *level* when p writes to *turn*. At some point q will notice that $turn \neq q$, and set back to 0, all the entries of the array V which equal to q. Repeat this argument with the new highest process. Thus, any entry of the array V which process p may wait on, will eventually be set back to 0, enabling p to proceed until it is elected. A contradiction. □

We say that a process is at level k, when the value of its private *level* register is k.

LEMMA 2.3. *For any* $k \in \{1, ..., \lceil \log n \rceil\}$*, out of all the processes that are in level* k *during a time interval where* $V[k]$ *continuously holds the value 0, at most one process can: (1) continue level* $k+1$ *or (2) change any register other than* $V[k]$*.*

PROOF. Assume that a set of processes $p_1, ..., p_\ell$ are at level k, and during the time interval where $V[k]$ continuously holds the value 0, they all notice that $V[level] = 0$ when executing the until statement in line 9. One of these processes, say p_1, must be the last to update *turn*. If $k = 1$, each process in $\{p_2, ..., p_\ell\}$ will notice that *turn* is different from its id (line 11), possibly write 0 into $V[1]$, and return 0. Assume $k > 1$. Before p_1 has set *turn* to its id, each of the other processes at level k, must have seen in level $k-1$ that *turn* is equal to its id. This means that before any of the processes $p_2, ..., p_\ell$ could execute the assignment at line 10, p_1 has already set $V[1], ..., V[k-1]$ to its id. Thus, when each process at level k, other than p_1, executes the if statement in line 11, it finds out that *turn* is different from its id,

possibly write 0 into $V[k]$, and returns 0, without a need to write 0 to any of the registers $V[1], ..., V[k-1]$. Process p_1, may continue to level $k+1$ or itself notices that $turn \neq p_1$ and sets some or all of the registers $V[1], ..., V[k-1]$ to 0, but its is the only process, among the processes $p_1, ..., p_\ell$, that may set any register other than $V[k]$. □

THEOREM 2.4 (SAFETY). *At most one leader is elected.*

For proving the theorem, an accounting system of credits is used. Initially, the number of credits is $2n-1$. New credits can not be created during the execution of the algorithm. The credit system ensures that a process acquires exactly 2^{k-1} credits before it can reach level k. Being elected is equivalent to reaching level $\log n+1$. Thus, the credit system ensures that a process must acquire $2^{\log n+1-1} = n$ credits before it can be elected. Once a process is elected, it may not release any of its credits. Thus, it is not possible for two processes to get elected.

With out loss of generality it is assumed that n, the number of processes, is a power of 2. Initially, each process holds 1 credit, and each register $V[k]$ where $1 \leq k \leq \log n$ holds 2^{k-1} credits. Thus, the total number of credits is $n + \sum_{k=1}^{\log n} 2^{k-1} = 2n-1$. As a results of an operation taken by a process credits may be transferred from a register to a process and vice versa. We list below all possible operations by processes and their effect:

- No credits are transferred when a process (1) checks the value of a register, (2) writes into *turn*, or (3) executes a *return* statement.
- When a process writes its id into register $V[k]$, changing $V[k]$'s value from 0 to its id, 2^{k-1} credits are transferred from $V[k]$ to that process. When a process writes 0 into register $V[k]$ which does not already holding 0, 2^{k-1} credits are transferred to $V[k]$ from that process.
- Let one or more processes notice that $V[k] = 0$. By Lemma 2.3, at most one of them can continue level $k+1$. Assume one of them continues to level $k+1$. By Lemma 2.3, the processes that do not continue to the next level can only execute $V[k] := 0$, transferring to $V[k]$ the 2^{k-1} credits they have by getting this far. Then 2^{k-1} credits are take from $V[k]$, and are assigned to the process that continues to the next level, giving it the 2^k credits it needs for level $k+1$.
- Let one or more processes notice that $V[k] = 0$, and assume no one of them continues to level $k+1$. By Lemma 2.3, at most one of these processes, say process p, changes any register other than $V[k]$. As before, the remaining processes can transfer their credits by setting $V[k]$ to 0. Then, if p is the last to set $V[k]$, 2^{k-1} credits are taken from $V[k]$, and are assigned to p. Thus, p has 2^k credits available, 2^{k-1} credits from reaching level k, plus 2^{k-1} credits from $V[k]$. Setting to 0 every variable from $V[1]$ to $V[k]$ accounts for $2^k - 1$ credits, so p has enough credits and no new credits should be created by p when it sets to 0 multiple registers.

As already mentioned, initially, the number of credits is $2n-1$. No new credits are created, and a process must acquire n credits before it can be elected. Once a process is elected, it may not release any of its credits. Thus, it is not possible for two processes to get elected. □

THEOREM 2.5 (ALMOST-WAIT-FREEDOM). *In the absence of faults, every participating process eventually terminates. In the presence of faults, every correct participating process, except maybe one, eventually terminates.*

PROOF. Once a leader is elected and returns, all correct participating processes will eventually find out that $done = 1$ and properly terminate. In particular, in the absence of faults, since by Theorem 2.2 at least leader is eventually elected, all the participating processes will terminate. Also, regardless of the number of faults, a correct process which is not the last to write into *turn*, will eventually either notices this fact and terminates or be elected and terminates. Thus, in the presence of faults, only the last process to write into *turn* may be blocked. □

2.3 Election in a message passing system

We present a simple election algorithm in which the process with the maximum identifer is elected.

THEOREM 2.6. *There is an almost-wait-free symmetric election algorithm with $n^2 - n$ message complexity.*

PROOF. In the algorithm each process sends its identifer to every other process, and collects, through the messages seen, identifiers of other processes. As soon as a process collects an identifer which is bigger than itself it returns 0. If a process collects the identifiers of all the other $n-1$ processes, and finds out that it is the process with the maximum identifer, it returns 1. In the code below *my.id* refers to the identifier of the process executing the algorithm, and *message.val* refers to the value of the message received. Each process has a local *counter* variable which is initially set to 0.

ALMOST-WAIT-FREE SYMMETRIC ELECTION ALGORITHM: program for a process with identifier *my.id*.

```
1 send my.id to all the other processes;
2 each time a message is received do
3    if my.id < message.val then
4       return(0) else counter := counter + 1 fi;
5    if counter = n − 1 then return(1) fi      /* leader! */
6 od
```

Clearly, in the absence of faults exactly one process is elected and it is always the process with the maximum identifier. In the presence of faults, only the correct participating process with the maximum identifier *among* the currently participating processes may not terminate, all the other processes will get a message from it, return 0 and terminate. The message complexity is $n^2 - n$, since each process sends one message to each other process. □

3. ALMOST-WAIT-FREE SYMMETRIC TEST-AND-SET BIT

We show that n registers are necessary and $n+1$ registers are sufficient for implementing a single almost-wait-free test-and-set bit using registers for n processes. A test-and-set bit supports two atomic operations called *test-and-set* and *reset*. A test-and-set operation takes as argument a shared bit b,

assigns the value 1 to b, and returns the previous value of b (which can be either 0 or 1). A reset operation takes as argument a shared bit b and writes the value 0 into b.

The *sequential specification* of an object specifies how the object behaves in sequential runs, that is, in runs when its operations are applied sequentially. The sequential specification of a test-and-set bit is quite simple. In sequential runs, the first test-and-set operation returns 0, a test-and-set operation that happens immediately after a reset operation also returns 0, and all other test-and-set operations return 1. We require that, although operations of processes may overlap, each operation should appear to take effect instantaneously. In particular, operations that do not overlap should take effect in their "real-time" order. This correctness requirement is called *linearizability* [24].

3.1 Upper bound

The algorithm below is for n processes each with a unique identifier taken from some (possibly infinite) set which does not include 0. It makes use of exactly n registers which are long enough to store a process identifier and one atomic bit. The algorithm is based on the symmetric mutual exclusion algorithm presented in [33].

THEOREM 3.1. *There is an almost-wait-free symmetric algorithm which implements a test-and-set bit using atomic registers. The algorithm is for n processes and uses $n+1$ atomic registers.*

The algorithm uses a register called *turn* to indicate who has priority to return 1, $n-1$ *lock* registers to ensure that at most one process will return 1 between resets, and a bit call *winner* to indicate whether some process already returned 1. Initially the values of all these shared registers are 0. In addition each process has a private boolean variable call *locked*. We denote by *b.turn*, *b.winner* and *b.lock*[*] the shared registers of the specific test-and-set *bit* named b.

AN ALMOST-WAIT-FREE SYMMETRIC TEST-AND-SET BIT:
`process p's program.`

```
function test-and-set (b:bit) return:value in {0,1};
1   if b.turn ≠ 0 then return(0) fi;                    /* lost */
2   b.turn := p;
3   repeat
4       for j := 1 to n − 1 do                          /* get locks */
5           if b.lock[j] = 0 then b.lock[j] := p fi od
6       locked := 1;
7       for j := 1 to n − 1 do                          /* have all locks? */
8           if b.lock[j] ≠ p then locked := 0 fi od;
9   until b.turn ≠ p or locked = 1 or b.winner = 1;
10  if b.turn ≠ p or b.winner = 1 then
11      for j := 1 to n − 1 do                          /* lost, release locks */
12          if b.lock[j] = p then b.lock[j] := 0 fi od
13      return(0) fi;
14  b.winner := 1; return(1).                           /* wins */
end_function

function reset (b:bit);                                 /* access bit b */
1   b.winner := 0; b.turn := 0;                         /* release locks */
2   for j := 1 to n − 1 do
3       if b.lock[j] = p then b.lock[j] := 0 fi od.
end_function
```

In the test-and-set operation, a process, say p, initially checks whether $b.turn = 0$, and if so returns 0. Otherwise, p takes priority by setting $b.turn$ to p, and attempts to obtain all the $n-1$ locks by setting them to p. This prevents other processes that also saw $b.turn = 0$ and set $b.turn$ to their ids from entering. That is, if p obtains all the locks before the other processes set $b.turn$, they will not be able to get any of the locks since the values of the locks are not 0. Otherwise, if p sees $b.turn \neq p$ or $b.winner = 1$, it will release the locks it holds, allowing some other process to proceed, and will return 0. In the reset operation, p sets $b.turn$ to 0, so the other processes can proceed, and releases all the locks it currently holds.

3.2 Correctness proof

We prove that our implementation is linearizable w.r.t. the sequential specification of a test-and-set bit mentioned earlier. For that it is enough to prove the following theorems. We say that run is *well structured*, if in that run a reset operation may be initiated only by a process that its last operation (before applying the reset operation) is a test-and-set operation which has returned 0. We say that a process is a *winner* in a given finite run, if the *last* completed operation of that process in the run is a test-and-set operation which has returned 0.

THEOREM 3.2 (SAFETY). *There is at most one winner in any well structured run.*

PROOF. Assume some process p is a winner. We show that no other process can become a winner before p preforms a reset operation. When process p last accessed *turn* and the $n-1$ locks, the value of each of these n shared registers was p. Any other process has to set all the $n-1$ locks and see *turn* set to its value for it to become a winner. But a process always checks a lock before writing it, and can only change one lock which has been already set (and not released yet) by some other process. So if all the n shared registers have the value p, and each of the remaining $n-1$ processes can overwrite at most one such register, at least one shared register must still hold the value p, preventing processes other than p from becoming winners. □

We say that a pending test-and-set operations is *potentially successful* if no process has become a winner since the operation was issued.

THEOREM 3.3 (LIVENESS). *In the absence of faults, at least one process will eventually become a winner, in any given run with potentially successful pending test-and-set operations.*

PROOF. Assume to the contrary that no process will become a winner. Since no process becomes a winner, *turn* is not set back to 0, and hence *turn* must eventually have a nonzero value, say p, and this value will not change thereafter. Every participating process other than p will eventually notice $turn = p$, it will release the locks it holds, will return 0 and thereafter will not update any other registers because *turn* is not zero. At this point, since process p always finds $turn = p$, nothing is preventing process p from getting all the locks and becoming a winner. A contradiction. □

THEOREM 3.4 (ALMOST-WAIT-FREEDOM). *In the absence of faults, every participating process (i.e, pending operation) eventually returns. In the presence of faults, every correct participating process, except maybe one, eventually returns.*

PROOF. Once some process becomes the winner and returns 1, as long as the winner does not initiate a reset operation, all correct participating processes will eventually find out that *done* = 1 and return 0. In particular, in the absence of faults, since by Theorem 3.3 at least one process will eventually become the winner, all the participating processes will return. Also, regardless of the number of faults, a correct process which is not the last to write *turn*, will eventually either notice this fact and return 0 or becomes the winner and returns 1. Thus, in the presence of faults, only the last process to write *turn* may be blocked. □

3.3 Lower bound

We show that the $n+1$ space upper bound is almost tight.

OBSERVATION 3.5. *Even in the absence of faults, any implementation of a test-and-set bit for n processes using atomic registers must use at least n atomic registers.*

PROOF. In [15, 16], it is proven that any deadlock-free mutual exclusion algorithm for n processes must use at least n shared registers. On the other hand, it is trivial to implement a deadlock-free mutual exclusion algorithm for n processes using a single test-and-set bit, say x, as follows: A process first keeps on accessing x until, in one atomic step, it succeeds to change x from 0 to 1. Then, the process can safely enter its critical section. The exit code is to reset x to 0. It is trivial to show that the algorithm satisfies mutual exclusion and is deadlock-free. The result follows. □

4. PARTIALLY-WAIT-FREE SYMMETRIC PERFECT RENAMING

A *renaming* algorithm allows processes with initially distinct initial names from a large name space to acquire distinct new names from a small name space. A *perfect* renaming algorithm allows n processes with initially distinct names from a large name space to acquire distinct new names from the set $\{1, ...n\}$. A *one-shot* renaming algorithm allows each process to acquire a distinct new name just once. A *long-lived* renaming algorithm allows processes to repeatedly acquire distinct names and release them (however, once a process acquires a new name it must first release it before trying to acquire another one).

It is well known that, in asynchronous systems where processes communicate either by reading and writing atomic registers or by sending and receiving messages, there is no 1-resilient perfect renaming algorithm [7, 30, 35]. Contrary to this impossibility result, we show that there is a partially-wait-free perfect renaming algorithm in such systems. A *partially-wait-free* renaming algorithm, should guarantee that t failures, where $1 \leq t \leq n-1$, may prevent at most t correct participating processes from acquiring new names.

THEOREM 4.1. *There is a partially-wait-free symmetric one-shot perfect renaming algorithm using either* (1) $n-1$ *almost-wait-free election objects,* (2) $O(n \log n)$ *registers, or* (3) $O(n^3)$ *messages.*

PROOF. First we present an algorithm which uses $n-1$ almost-wait-free election objects. The election objects are indexed 1,2,....,$n-1$. Each process scans the objects, in order, starting with object number 1. At each step, the process applies the election operation, and either: moves to the next object if it is not elected in object $i < n-1$, stops if it is being elected, or stops if it not elected in object $n-1$. The process is assigned either the name equal to the index of the object on which its election operation has succeeded, or n if it is not elected in all $n-1$ objects. Notice that at most $n-i+1$ processes may participate in object i, for $1 \leq i \leq n-1$. Thus, by Theorem 2.1, the almost-wait-free, election object indexed i, where $1 \leq i \leq n-1$, can be implemented using $\lceil \log(n-i+1) \rceil + 2$ atomic registers. Thus, the number of registers used are at most:

$$3(n-1) + \sum_{i=2}^{n} \log i = 3(n-1) + \log n! = O(n \log n).$$

Finally, by Theorem 2.6, there is an implementation of an almost-wait-free symmetric election object for n processes which has n^2-n message complexity. The result follows. □

THEOREM 4.2. *There is a partially-wait-free symmetric long-lived perfect renaming algorithm using either* $n-1$ *almost-wait-free test-and-set bits* or $O(n^2)$ *atomic registers.*

PROOF. First we present an algorithm which uses $n-1$ almost-wait-free test-and-set bit bits. The bits have initial values 0, and are indexed 1,2,....,$n-1$. Each process scans the bits, in order, starting with bit number 1. At each step, the process applies a *test-and-set* operation, and either: moves to the next bit if the returned value is 1 in bit $i < n-1$, stops when the returned value is 0, or stops if the returned value is 1 in bit $n-1$. The process is assigned the name equal to the index of the bit on which its (last) *test-and-set* operation returned 0, or n if the returned value is 1 in all $n-1$ bits. A process which is assigned the name i can later release this name by applying a *reset* operation to the i'th bit setting its value back to 0. A process which is assigned the name n doesn't have to access any shared bit to release the name n. At most $n-i+1$ processes may concurrently access the bit indexed i, for $1 \leq i \leq n-1$. Thus, by Theorem 3.1, the bit indexed i, where $1 \leq i \leq n-1$, can be implemented using $n-i+2$ registers. Thus, the number of registers used are:

$$\sum_{i=2}^{n}(i+1) = \frac{n^2+3n-4}{2}.$$

The result follows. □

5. PARTIALLY-WAIT-FREE FETCH-AND-ADD, SWAP, AND STACK

A *fetch-and-add* object supports one operation, which takes as arguments a shared register r, and a value *val*. The value of r is incremented by *val*, and the old value of r is returned. A *swap* object supports one operation, which takes as arguments a shared registers and a local register and atomically exchange their values. A concurrent *stack* is a linearizable object that supports push and pop operations, by several processes, with the usual stack semantics. A *sequential* process is a process that has at most one pending operation at any given time.

LEMMA 5.1. *Assume that there is a wait-free implementation for n sequential processes of an object o using wait-free test-and-set bits and atomic registers. Then, there is a partially-wait-free implementation for n processes of o using atomic registers only.*

PROOF. Let A be a wait-free implementation for n sequential processes of an object o using wait-free test-and-set bits and registers. Let A' be the implementation A where each wait-free test-and-set bit is replaced with an almost-wait-free test-and-set bit. While executing A', a failure of a process with a pending test-and-set operation, may prevent at most one other process from completing its operation in A'. Thus, a failure of t processes may prevent at most t other process from completing their operations. This implies that A' is a partially-wait-free implementation of o using almost-wait-free test-and-set bits and registers. By Theorem 3.1, we can replace each almost-wait-free test-and-set bit in A', by an implementation using atomic registers. The result follows. □

THEOREM 5.2. *There are partially-wait-free implementations for n processes of a fetch-and-add object, a swap object, and a stack object using atomic registers only.*

PROOF. In [6], a class of shared objects called Common2 were defined. Each object in Common2 is known to have a wait-free implementation from registers together with any other object in Common2, for an arbitrary number of sequential processes. Commonly used primitives such as test-and-set, fetch-and-add, swap, and stack are in Common2 [2, 6]. Thus, any of the objects in Commom2 has a wait-free implemention using registers and wait-free test-and-set bits, for arbitrary number of sequential processes. (The implementations presented in [6] are not symmetric.) This last observation together with Lemma 5.1 implies that there are partially-wait-free implementations for n processes of a fetch-and-add object, a swap object, and a stack object using atomic registers only. □

6. IMPOSSIBILITY RESULTS FOR CONSENSUS AND SET-CONSENSUS

The *k-set consensus* problem is to find a solution for n processes, where each process starts with an input value from some domain, and must choose some participating process' input as its output. All n processes together may choose no more than k distinct output values. The 1-set consensus problem, is the familiar consensus problem for n processes.

The consensus and set-consensus problems belong to a class of problems called *colorless tasks*. Colorless tasks (also called convergence tasks [13]) allow a process to adopt an input or output value of any other participating process, so the task can be defined in terms of input and output sets instead of vectors.

For proving the following lemma we need to assume a model where participation is required. Recall that with required participation every process must eventually execute its code.

LEMMA 6.1. *Assume a model where participation is required, $n \geq 3$ and $t \leq n-2$. When processes communicate either by reading and writing atomic registers or by sending and receiving messages, for any colorless task T: there is a weakly-t-resilient algorithm which solves T if and only if there is a t-resilient algorithm which solves T.*

PROOF. Let A be a weakly-t-resilient algorithm using atomic registers which solves T. We use A to implement a t-resilient algorithm, called A', which uses atomic registers and solves T. An additional shared register called *output* is used, which has initial value $\perp$. Every process executes as in A, and before it terminates it writes its output into *output*. During its execution of A, a process also continuously checks whether *output* $\neq \perp$, and in case the test is positive, it adopts the value of *output* as its own output value and terminates. Since participation is required, $n \geq 3$ and $t \leq n-2$, one correct process will eventually terminate. Once one correct process writes its output into *output*, it is guaranteed that each participating correct will eventually either terminates according its code in A, or will notice that *output* $\neq \perp$, and properly terminates. The resulting algorithm is A'. Proving the other direction is trivial. The proof for the case where communication is by sending and receiving messages is almost the same. Instead of writing to *output*, a process sends its decision to everyone before terminating. Each process that receives a message with such a decision value, decides on that value, sends it to everyone and terminates. □

The following results hold for a model where participation is required, and thus also hold for a model where participation is not required.

THEOREM 6.2. *For $n \geq 3$, there is no weakly-1-resilient consensus algorithm using either reading and writing atomic registers or sending and receiving messages.*

PROOF. The proof follows from Lemma 6.1 and the known result that there is no 1-resilient consensus algorithm using either reading and writing atomic registers or sending and receiving messages [20, 28]. This known impossibility result was proved for a model where participation is required and thus also trivially holds for a model where participation is not required. □

THEOREM 6.3. *For $n \geq 3$ and $1 \leq k \leq n-2$, there is no weakly-k-resilient k-set-consensus algorithm using either reading and writing atomic registers or sending and receiving messages.*

PROOF. The proof follows from Lemma 6.1 and the known result the there for $1 \leq k \leq n-1$, is no k-resilient k-set-consensus algorithm for n processes using atomic registers [12, 23, 32]. The impossibility result for the message passing model follows immediately from the one for the shared memory model. This known impossibility result was proved for a model where participation is required and thus also trivially holds for a model where participation is not required. □

COROLLARY 6.4. *For $n \geq 3$ and $1 \leq k \leq n-2$, there is no weakly-k-resilient k-set-consensus algorithm using almost-wait-free test-and-set bits and atomic registers.*

PROOF. The proof follows immediately from Theorem 3.1 and Theorem 6.3. □

7. DISCUSSION

We have refined the traditional notion of *t-resiliency* by defining the finer grained notion of *(t, f)-resiliency*. In particular, we have extended the investigation of fault-tolerance by presenting several new notions: weakly-t-resiliency, partially-t-resiliency and almost-t-resiliency.

In the traditional notion of *t-resiliency* it is assumed that failures are *uniform*: processes are equally probable to fail,

and failure of one process does not affect the reliability of the other processes. As discussed in [27], in real systems, failures may be correlated because of software or hardware features shared by subsets of processes. Our new resiliency notions can be defined similarly also for such *non-uniform* failure models, and it would be interesting to extend our results to cover such failure models.

All our results are presented in the context of crash failures in asynchronous systems, it would be interesting to consider also other types of failures such as omission failures and Byzantine failures, and to consider synchronous systems. Another interesting direction would be to extend the results for other objects. In particular, is there an almost-wait-free (or even a weakly-wait-free) implementation of a shared *queue* object from registers? We have assumed that the number of processes is finite and known, it would be interesting to consider also the case of unbounded concurrency. Considering failure detectors in the context of the new definition is another interesting direction.

Several other questions are left open. We have presented a symmetric almost-wait-free implementation of a test-and-set bit from registers. Are there similar symmetric almost-wait-free implementations for, stack, swap and fetch-and-add objects from registers? In case that there is no almost-wait-free perfect renaming, what is the smallest m for which there is a solution for almost-wait-free renaming in which a process always gets a distinct name in the range $\{1, ..., m\}$? Finally, are there implementations which are more space, time or message efficient than the implementations presented?

8. REFERENCES

[1] Y. Afek, H. Attiya, A. Fouren, G. Stupp, and D. Touitou. Long-lived renaming made adaptive. In *Proc. 18th ACM Symp. on Principles of Distributed Computing*, pages 91–103, May 1999.

[2] Y. Afek, E. Gafni, and A. Morrison. Common2 extended to stacks and unbounded concurrency. In *Proc. 25th ACM Symp. on Principles of Distributed Computing*, pages 218–227, 2006.

[3] Y. Afek and M. Merritt. Fast, wait-free $(2k-1)$-renaming. In *Proc. 18th ACM Symp. on Principles of Distributed Computing*, 105–112, 1999.

[4] Y. Afek, G. Stupp, and D. Touitou. Long-lived adaptive collect with applications. In *Proc. 40th IEEE Symp. on Foundations of Computer Science*, pages 262–272, Oct. 1999.

[5] Y. Afek, G. Stupp, and D. Touitou. Long lived adaptive splitter and applications. *Distributed Computing*, 30:67–86, 2002.

[6] Y. Afek, E. Weisberger, and H. Weisman. A completeness theorem for a class of synchronization objects (extended abstract). In *Proc. 12th ACM Symp. on Principles of Distributed Computing*, pages 159–170, 1993.

[7] H. Attiya, A. Bar-Noy, D. Dolev, D. Koller, D. Peleg, and R. Reischuk. Achievable cases in an asynchronous environment. In *Proc. 28th IEEE Symp. on Foundations of Computer Science*, 337–346, Oct. 1987.

[8] H. Attiya, A. Bar-Noy, D. Dolev, D. Koller, D. Peleg, and R. Reischuk. Renaming in an asynchronous environment. *Journal of the Association for Computing Machinery*, 37(3):524–548, July 1990.

[9] H. Attiya and A. Fouren. Polynomial and adaptive long-lived $(2k-1)$-renaming. In *Proc. 14th International Symp. on Distributed Computing: Lecture Notes in Computer Science 1914*, pages 149–163, Oct. 2000.

[10] H. Attiya and A. Fouren. Algorithms adapting to point contention. *Journal of the ACM*, 50(4):144–468, 2003.

[11] A. Bar-Noy and D. Dolev. Shared memory versus message-passing in an asynchronous. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pages 307–318, 1989.

[12] E. Borowsky and E. Gafni. Generalizecl FLP impossibility result for t-resilient asynchronous computations. In *Proc. 25th ACM Symp. on Theory of Computing*, pages 91–100, 1993.

[13] E. Borowsky, E. Gafni, N. A. Lynch, and S. Rajsbaum. The BG distributed simulation algorithm. *Distributed Computing*, 14(3):127–146, 2001.

[14] A. Brodsky, F. Ellen, and P. Woelfel. Fully-adaptive algorithms for long-lived renaming. *Distributed Computing*, 24(2):119–134, 2011.

[15] J. Burns and A. Lynch. Mutual exclusion using indivisible reads and writes. In *18th annual allerton conference on communication, control and computing*, pages 833–842, Oct. 1980.

[16] J. Burns and N. Lynch. Bounds on shared-memory for mutual exclusion. *Information and Computation*, 107(2):171–184, Dec. 1993.

[17] J. Burns and G. Peterson. The ambiguity of choosing. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pages 145–158, Aug. 1989.

[18] A. Castaneda, S. Rajsbaum, and M. Raynal. The renaming problem in shared memory systems: An introduction. *Computer Science Review*, 5(3):229–251, 2011.

[19] C. Delporte-Gallet, H. Fauconnier, R. Guerraoui, and A. Tielmanns. The disagreement power of an adversary. In *Proc. 28th ACM Symp. on Principles of Distributed Computing*, pages 288–289, 2009.

[20] M. Fischer, N. Lynch, and M. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, 1985.

[21] E. Gafni, M. Merritt, and G. Taubenfeld. The concurrency hierarchy, and algorithms for unbounded concurrency. In *Proc. 20th ACM Symp. on Principles of Distributed Computing*, pages 161–169, Aug. 2001.

[22] M. P. Herlihy. Wait-free synchronization. *ACM Trans. on Programming Languages and Systems*, 13(1):124–149, Jan. 1991.

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

[24] M. P. Herlihy and J. M. Wing. Linearizability: a correctness condition for concurrent objects. *toplas*, 12(3):463–492, 1990.

[25] D. Imbs, M. Raynal, and G. Taubenfeld. On asymmetric progress conditions. In *Proc. 29th ACM Symp. on Principles of Distributed Computing*, pages 55–64, 2010.

[26] M. Inoue, S. Umetani, T. Masuzawa, and H. Fujiwara.

Adaptive long-lived $O(k^2)$-renaming with $O(k^2)$ steps. In *15th international symposium on distributed computing*, 2001. *LNCS 2180* Springer Verlag 2001, 123–135.

[27] P. Kuznetsov. Understanding non-uniform failure models. *Distributed computing column of the Bulletin of the European Association for Theoretical Computer Science (BEATCS)*, 106:54–77, 2012.

[28] M. Loui and H. Abu-Amara. Memory requirements for agreement among unreliable asynchronous processes. *Advances in Computing Research*, 4:163–183, 1987.

[29] M. Moir and J. H. Anderson. Wait-free algorithms for fast, long-lived renaming. *Science of Computer Programming*, 25(1):1–39, Oct. 1995.

[30] S. Moran and Y. Wolfstahl. Extended impossibility results for asynchronous complete networks. *Information Processing Letters*, 26(3):145–151, 1987.

[31] M. Pease, R. Shostak, and L. Lamport. Reaching agreement in the presence of faults. *Journal of the ACM*, 27(2):228–234, 1980.

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

[33] E. Styer and G. L. Peterson. Tight bounds for shared memory symmetric mutual exclusion problems. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pages 177–191, Aug. 1989.

[34] G. Taubenfeld. The computational structure of progress conditions. In *24th international symposium on distributed computing (DISC 2010)*, Sept. 2010. *LNCS 6343* Springer Verlag 2010, 221–235.

[35] G. Taubenfeld and S. Moran. Possibility and impossibility results in a shared memory environment. *Acta Informatica*, 33(1):1–20, 1996.