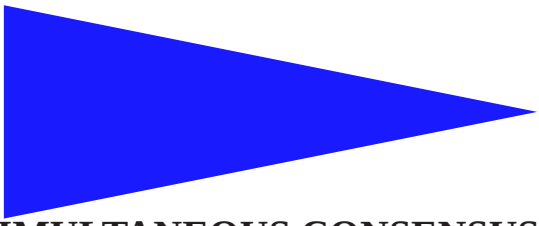
The k-simultaneous consensus problem

Yehuda Afek, Eli Gafni, Sergio Rajsbaum, Michel Raynal, Corentin Travers

PUBLICATION
INTERNE
N° 1920



THE K-SIMULTANEOUS CONSENSUS PROBLEM

Y. AFEK E. GAFNI S. RAJSBAUM M. RAYNAL C. TRAVERS

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

The k-simultaneous consensus problem

Y. Afek[*] E. Gafni[**] S. Rajsbaum[***] M. Raynal[****] C. Travers[*****]

Systèmes communicants
Projet ASAP

Publication interne n ° 1920 — Janvier 2009 — 15 pages

Abstract: This paper introduces and investigates the k-simultaneous consensus problem: each process participates at the same time in k independent consensus instances until it decides in any one of them. Two results are presented. The first shows that the k-simultaneous consensus problem and the k-set agreement problem are wait-free equivalent in read/write shared memory systems. The second shows that the multi-valued version and the binary version of the k-simultaneous consensus problem are wait-free equivalent. These equivalences are independent of the number of processes. An immediate consequence of these results is that the k-set agreement problem and the k-simultaneous binary consensus problem are equivalent. This not only provides a new characterization of the k-set agreement problem but also provides a meaning to the notion of k-set binary agreement.

Key-words: Asynchronous shared memory systems, Binary vs multivalued agreement, Consensus, Distributed computability, Process crash, Set agreement, Wait-free construction.

(Résumé : tsvp)

[*] Computer Science Department, Tel-Aviv University, Israel 69978. afek@math.tau.ac.il.
[**] Department of Computer Science, UCLA, Los Angeles, CA 90095, USA, eli@cs.ucla.edu.
[***] Instituto de Matemáticas, UNAM, D. F. 04510, Mexico. rajsbaum@math.unam.mx.
[****] Université de Rennes 1, IRISA, Campus de Beaulieu, 35042 Rennes, France, raynal@irisa.fr.
[*****] Computer Science Department, The Technion, Haifa, Israel, travers@gmail.fr.

A preliminary draft of this paper has been presented at the conference ICDCN'06 [2].

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Le problème du consensus simultané avec k instances

Résumé : Ce rapport introduit et analyse le problème du consensus simultané avec k instances.

Mots clés : Accord ensembliste, Calculabilité distribuée, Consensus binaire, Consensus multivalué, Crash de processus, Mémoire partagée, Synchronisation sans attente, Système asynchrone.

1 Introduction

Context and motivation of the paper The consensus problem is one of the most basic coordination problems in distributed systems. Assuming each process proposes a value, it requires that (1) each non-faulty process decides on a value (*termination*) in such a way that (2) there is a single decided value (*agreement*), and (3) the decided value is one of the proposed values (*validity*). Unfortunately, this problem has no solution in asynchronous systems as soon as even only one process may crash, be the system a shared memory system [18] or a message passing system [10].

One way to weaken the consensus problem is to allow several different values to be decided. This approach has given rise to the *k-set agreement* problem where up to k different values can be decided [7]. While this problem (sometimes also called k-set consensus) can be solved despite asynchrony and process failures when $k > t$ (where t is the maximum number of processes that can be faulty), it has been shown that it has no solution when $t \geq k$ [5, 15, 22].

This paper presents and investigates another way to weaken the consensus problem. The intuition that underlies this problem, called here *scalar k-simultaneous consensus*, is "win one out of several". More explicitly, each process proposes a value to k independent consensus instances (namely, the same value to all these instances). It is required that every correct process decides on a value in only one consensus instance. So, a process decides on a pair composed of a value and a consensus instance number. Two processes can decide on two different pairs; however if they decide on the same consensus instance they also decide on the same value (that has been proposed by one of the processes)[1]. We also consider an equivalent *vector* version of the *k-simultaneous consensus*, where, each process proposes k possibly different values, one value to each of the k independent consensus instances. Again a process decides on a pair composed of a value and a consensus instance number. Two processes can decide on two different pairs; if they decide on the same consensus instance they also decide on the same value (that has been proposed to that instance). It is easy to see that the two versions of the problem, the scalar version and the vector version, are equivalent (see Section 2).

As explained in [13], simultaneous consensus can be useful in situations where several processes participate concurrently in k different applications: a k-simultaneous consensus solution can guarantee wait-free progress in at least one application[2]. Indeed, recently this problem has been instrumental in determining the weakest failure detector for wait-free solving the $(n-1)$-set agreement problem in asynchronous read/write shared memory systems made up of n processes [23] (wait-free means that the solution has to cope with any number of process failures [14]). More generally, the simultaneous consensus problem captures the situations where a process has to decide in any one of several consensus instances. Moreover, in addition to its possible applications, various interesting generalizations are fairly natural, namely, each consensus instance can be instead some other agreement problem, for example a ℓ-set agreement instance (the Conclusion section discusses this specific case).

Content of the paper This paper originated from the following questions.

- The first question addresses the relation between the k-set agreement problem and the k-simultaneous consensus problem. While, given a solution to the k-simultaneous consensus problem, it is easy to solve the k-set agreement problem, what about the other direction? Said in another way, are these problems equivalent or not?

[1] Let us notice that the words "simultaneous consensus" have been used with a different meaning in round-based synchronous systems. In these systems, they mean that all the processes that participate in a consensus instance have to terminate during the very same round [8, 9].

[2] The k-simultaneous problem is close but different from the BG simulation [5, 6]. Here each process systematically participates in k concurrent consensus. In the BG simulation, a process participates in a new agreement only if it is blocked in the previous one.

The paper answers this question positively by showing that they are indeed equivalent. To that end, it presents a wait-free transformation that, given a k-set agreement object, builds a k-simultaneous consensus object.

- The second question concerns the relation between the k-simultaneous multivalued consensus and its binary counterpart. More precisely, while it is known that multivalued consensus and binary consensus are equivalent (e.g., [21]), what about k-simultaneous multivalued consensus and k-simultaneous binary consensus? The binary version of the problem is a simple case of the multivalued one, but what in the other direction?

  The paper shows that the two problems are equivalent. To that end, it presents a wait-free transformation, that, given k-simultaneous binary consensus objects, builds a k-simultaneous multivalued object.

Hence, the paper shows that the k-set agreement problem and the k-simultaneous binary consensus problem are equivalent. Thus, while k-set agreement, unlike consensus, has no binary version, the previous equivalence provides a characterization of k-set agreement in terms of k-simultaneous binary consensus.

The origin of the problem The line of research developed in this paper originates from [12, 13]. More precisely, first [12] investigated the *musical benches* problem (inspired by the musical chairs children's game) to model a new distributed coordination difficulty, namely, processes jump from bench to bench trying to find one in which they may be alone or not in conflict with one another (of course, there is one more process than benches, and each bench is made up of two places). Then, a binary version of the vector consensus problem was introduced in [13] under the name *committee decision* problem (each committee being a consensus instance). It is shown in [13] that, with binary inputs an for three processes, musical benches, 2-set agreement and 2-simultaneous consensus are wait-free equivalent.

The results in [12, 13] have been obtained through a novel reduction technique that combines distributed algorithmic ideas with topological intuition, but are for three processes only. The equivalence of set agreement and simultaneous consensus was extended to any number of processes in [2] (which is the conference version of this paper).

Roadmap Section 2 describes the computation model and presents the base problems we are interested in. Then, Section 3 shows that the k-set agreement problem and the k-simultaneous consensus problem are equivalent. Section 4 shows that the k-simultaneous multivalued consensus problem its not more powerful than its binary counterpart. Finally, Section 5 concludes the paper.

2 Computation model and problem definitions

2.1 Computation model

Processes The system consists of an arbitrary number of processes denoted $p_i, p_j, \ldots$ [11, 20]. The integer i is the identity of p_i, and no two processes have the same identity. Each run is characterized by an unknown number of *participating* processes. A process that participates in a run knows neither the identities of the processes that participate in that run, nor their number.

There is no assumption on the speed of processes: they are asynchronous. Moreover, any number of processes may crash. Before it crashes (if it ever crashes), a process executes correctly its algorithm. A crash is a premature halt: after it has crashed, a process executes no more statement. Given a run, a process that does not crash is *correct* in that run, otherwise it is *faulty* in that run.

A remark on the number of processes Most distributed algorithms are designed for a set of n processes where n is fixed and known by every process. Moreover, each process is assigned a unique identity comprised between 1 and n, and an algorithm can make use of both the number of processes and their identity.

Differently, the algorithms designed in this paper work with an arbitrary number of processes. Such a situation occurs in systems that dynamically change over time. For example, a network may allow nodes to be added or removed, or an operating system may allow processes to dynamically join, participate in a distributed algorithm and finally leave. Algorithms for infinitely many processes (e.g. [11, 20]) have recently received attention. Their advantages over algorithms for a fixed number of processes are significant [3]: (1) They have no system size parameters to configure, and (as a result) they are more robust and elegant; (2) They automatically handle crash recovery of processes (as a process that crashes and recovers can join the algorithm simply by assuming a new identity); (3) They guarantee progress even if processes keep on arriving (which is important in loosely-coupled systems, like peer-to-peer systems, where there is a large number of nodes that come and go all the time).

Communication model The processes communicate through reliable multi-reader/multi-writer atomic registers [4, 17, 19]. In addition the algorithms presented here make heavy use of the atomic snapshot primitive [1]. This basic operation, denoted $\mathsf{snapshot\_scan}(A)$ where A is an array in the shared memory with one entry per process, returns a copy of all the values in the array such that they were simultaneously present in A during the snapshot operation. Therefore, if each cell in the array is written only once and values cannot be removed, the sequence of snapshots obtained by repetitive $\mathsf{snapshot\_scan}(A)$'s is a sequence of growing snapshots such that each snapshot includes all the previous snapshots in the sequence: the $\mathsf{snapshot\_scan}()$ invocations are linearizable [16] (an we say that the sequence of growing snapshots satisfies the *containment* property). A wait-free snapshot algorithm that works in the case of unbounded concurrency is described in [11].

Input values In the agreement problems we investigate, the set of input values, denote $\mathcal{I}$, is totally ordered. If $\mathcal{I}$ is finite, n denotes its number of elements. If $\mathcal{I}$ contains only two values, the agreement problem is *binary*, otherwise it is *multivalued*. The value $\perp$ is a default value that does not belong to $\mathcal{I}$, and for any non-empty subset S of $\mathcal{I}$ we have: $\min(S \cup \{\perp\}) = \min(S)$ and $\max(S \cup \{\perp\}) = \max(S)$ (moreover, $\min(\{\perp\}) = \max(\{\perp\}) = \perp$).

Remark All the algorithms described in the paper are given for an arbitrary process p_i. Uppercase letters are used to denote shared objects, while lowercase letters are used for local variables (these variables are subscribed with the index of the corresponding process).

2.2 Problem definitions

The k-set agreement problem As indicated in the Introduction, the k-set agreement problem [7] is a generalization of the consensus problem (that corresponds to the case $k = 1$). It is defined by the following properties.

- Termination: each correct process decides on a value.
- Validity: a decided value is a proposed value.
- Agreement: at most k different values are decided.

As for all the problems considered in this paper, the termination property requires a solution based on wait-free algorithms [14]: a correct process has to terminate regardless of the number of faulty processes.

It is important to notice that, while the binary consensus problem is meaningful, the binary k-set agreement problem is meaningless when $k > 1$.

Let KSA be an object that solves the k-set agreement problem. It provides the processes with a single operation denoted $KSA.\mathsf{set\_propose}_k()$. That operation takes a proposed value as input parameter, and returns a decided value.

The k-simultaneous consensus problem Both (the scalar and the vector) versions of the k-simultaneous consensus problem consists of k independent instances of the consensus problem where a process is required to decide in one of them. (The 1-simultaneous consensus problem boils down to the consensus problem.) More precisely, in the *scalar version*, process p_i proposes the same value v_i to each of the consensus instances. In the *vector version*, process p_i proposes a vector $[v_i^1, \cdots, v_i^k]$ where v_i^c is the value it proposes to the c's consensus instance ($1 \leq c \leq k$). Each process decides on a pair (c, d) where c is a consensus instance and d is a value. The problem is defined by the following properties.

- Termination: each correct process decides on a pair.
- Validity: if a process p_i decides (c, d), then c is a consensus instance (i.e., $1 \leq c \leq k$), and d is a value that has been proposed to that consensus instance.
- Agreement: if two processes decide (c, d) and (c, d'), then $d = d'$.

Let KSC be an object that solves the k-simultaneous consensus problem. It provides the processes with a single operation denoted $KSC.\mathsf{sc\_propose}_k()$. In the scalar version that operation takes as input parameter the process input value, and in the vector version it takes a vector with k proposed values (one for each consensus instance), and returns a pair (c, d). In the case of a binary k-simultaneous consensus object, the operation is denoted $\mathsf{bin\_sc\_propose}_k()$ and all the input values are binary.

2.3 Problems equivalence

For comparing two problems we use wait-free constructions. Namely, for problems A, B, we say that "A *can implement* B" if there is a wait-free algorithm C that has access to any number of copies of A objects (in addition to read/write atomic registers) and solves the problem B; we say the algorithm C implements a B object. If A and B implement each other, the problems are said to be *equivalent*. All the constructions described in the paper are wait-free and work for an arbitrary number of processes.

2.4 The scalar version and the vector version are equivalent

It is easy to see that the vector version and the scalar version are equivalent, i.e., can implement each other. To implement the scalar version from the vector version process p_i proposes a size k vector $[v_i, \ldots, v_i]$. The algorithm described in Figure 1 implements the vector version from the scalar version. A process p_i executes sequentially the following statements: (1) it first publishes its vector in the shared memory (represented by an array denoted $INPUT$ with one entry per process); (2) then, p_i proposes its identity i to the underlying scalar version of the k-simultaneous consensus problem and obtains a pair (c_i, j) finally, (3) p_i decides on the pair (c, res) where res is the value in $INPUT[j][c_i]$ (i.e., the value proposed by p_j to the c_i-th consensus instance). The prof is easy and left to the reader.

3 k-Set agreement vs k-simultaneous consensus

This section shows that the k-set agreement problem and the scalar k-simultaneous consensus problem are equivalent. To that end it presents two wait-free constructions, one in each direction. Both constructions are independent on the number of processes.

```
operation KSC.sc_propose_k(v_i^1, ..., v_i^k): % vector version %
(01) INPUT[i] ← [v_i^1, ..., v_i^k];
(02) < c_i, j > ← KSC.sc_propose_k(i); % scalar version %
(03) let d_i = INPUT[j][c_i];
(04) return(c_i, d_i).
```

Figure 1: k-Simultaneous consensus: from the scalar version to the vector version

3.1 From scalar k-simultaneous consensus to k-set agreement

A pretty simple wait-free algorithm that builds a k-set agreement object (denoted KSA) on top of a k-simultaneous consensus object (denoted KSC) is described in Figure 2. The invoking process p_i calls the underlying object KSC with its input to the k-set agreement as input, and obtains a pair (c_i, d_i). It then returns d_i as the decision value for its invocation of $KSA.\mathsf{set\_propose}_k(v_i)$.

```
operation KSA.set_propose_k(v_i):
(01) (c_i, d_i) ← KSC.sc_propose_k(v_i);
(02) return(d_i).
```

Figure 2: From scalar k-simultaneous consensus to k-set agreement

Lemma 1 *The algorithm described in Figure 2 is a wait-free construction of a k-set agreement object from a scalar k-simultaneous consensus object.*

Proof The proof is immediate. The termination and validity of the k-set agreement object follow directly from the code and the same properties of the underlying k-simultaneous consensus object. The agreement property follows from the fact that at most k values can be decided from the k consensus instances of the k-simultaneous consensus object. $\square_{Lemma\ 1}$

3.2 From k-set agreement to scalar k-simultaneous consensus

A wait-free algorithm that constructs a scalar version k-simultaneous consensus object KSC from a k-set agreement object KSA is described in Figure 3.

```
operation KSC.sc_propose_k(v_i):
(01) dv_i ← KSA.set_propose_k(v_i);
(02) SM[i] ← dv_i;
(03) snap_i ← snapshot_scan(SM);
(04) let c_i = |snap_i|; let d_i = minimum value in snap_i;
(05) return(c_i, d_i).
```

Figure 3: From k-set agreement to scalar k-simultaneous consensus

In the algorithm, the processes first go through a k-set agreement object to reduce the number of distinct values to at most k (line 01). Then, each process p_i (1) posts the value it has just obtained in the cell $SM[i]$

of the shared memory (initialized to $\perp$), and (2) takes a snapshot of the whole shared memory (line 03). Finally, a process p_i returns the pair (c_i, d_i) where the consensus instance c_i is defined as the number of distinct values in the snapshot returned to p_i, and d_i is the minimum value in that snapshot.

Lemma 2 *The algorithm described in Figure 3 is a wait-free construction of a scalar k-simultaneous consensus object from a k-set agreement object.*

Proof The code in Figure 3 is wait-free since there are no loops and both the k-set agreement and the snapshot operations are wait-free. The validity follows from the fact that all the values in the algorithm originate from process inputs.

Since the snapshots by the different processes define a linearizable sequence ordered by containment, they also define a growing sequence when we consider the size of the snapshots returned to the processes. Therefore, there is a unique set of values contained in each snapshot size and hence the minimum value in each snapshot size is unique. Thus there are at most k distinct snapshot sizes, each with its unique minimum value. Hence, there are at most k distinct outputs returned and any two processes that return a pair with the same snapshot size (same first coordinate) have the same value associated with it, which proves the agreement property of the k-simultaneous consensus. $\square_{Lemma\ 2}$

3.3 A first equivalence

Theorem 1 *The k-set agreement problem and the scalar k-simultaneous consensus problem are wait-free equivalent in read/write shared memory systems made up of an arbitrary number of processes. Moreover, this equivalence is independent of the number of values that can be proposed.*

Proof The proof of the equivalence follows directly from Lemmas 1 and 2. The fact that this equivalence is independent of n follows from a simple examination of the text of the algorithms, where n never appears. $\square_{Theorem\ 1}$

4 Binary vs multivalued k-simultaneous consensus

The operation $\mathsf{bin\_sc\_propose}_k()$ is trivially a particular instance of the $\mathsf{sc\_propose}_k()$ operation: it corresponds to the case where only two values can be proposed ($|\mathcal{I}| = 2$). This section focuses on the transformation in the other direction. Assuming $|\mathcal{I}|$ is bounded and $n = |\mathcal{I}|$ is known to the processes, this section describes an algorithm that implements the scalar multivalued $\mathsf{sc\_propose}_k()$ operation from atomic registers and binary vector simultaneous consensus objects. Let us observe that, while every process knows n, none of them knows initially the values that define the set $\mathcal{I}$ (but the value it proposes).

4.1 A modular construction

The main construction presented in the next subsection builds an intermediary object, that we call a *restricted ℓ-simultaneous consensus* object. The aim of such an object is to reduce by one the number of proposed values. More precisely, assuming that at most $\ell + 1$ different values are proposed by the processes, this object guarantees that (1) each process decides a value, and (2) at most ℓ different values are decided on. The next subsection (Section 4.2) shows how a restricted ℓ-simultaneous consensus object can be built out of atomic registers and a binary vector ℓ-simultaneous consensus object.

Here we show how a cascading sequence of restricted ℓ-simultaneous consensus objects for $\ell = n - 1, n - 2, \ldots, k$ is used to construct a k-simultaneous consensus object KSC. Each restricted simultaneous

consensus object in the sequence reduces the number of different values by one and the whole sequence reduces $|\mathcal{I}|$ from n to k as described in Figure 4. Notice that a binary ℓ-simultaneous consensus is trivially implemented from binary k-simultaneous consensus for $\ell \geq k$, thus, all together we construct a multivalued k-simultaneous consensus from binary k-simultaneous consensus.

```
operation KSC.sc_propose_k(v_i):
(01) prop_i ← v_i;
(02) for ℓ from n − 1 step −1 to k do
(03)     (c_i, prop_i) ← RSC[ℓ].rsc_propose_ℓ(prop_i)
(04) end for;
(05) return(c_i, prop_i).
```

Figure 4: From restricted simultaneous consensus to scalar k-simultaneous consensus

Lemma 3 *The algorithm described in Figure 4 is a wait-free construction of a scalar k-simultaneous consensus object from restricted ℓ-simultaneous consensus objects, with $\ell = n-1, \cdots, k$.*

Proof The proof relies on the fact that the loop is made up of consecutive rounds. As there are initially at most n different values proposed by the processes, it follows from the definition of the $RSC[n-1]$ object that at most $n-1$ of these values are returned by the invocations $RSC[n-1].\mathsf{rsc\_propose}_{n-1}()$ issued by the processes. Then, the next rounds reduce the number of values to (at most) k. Finally, it follows from the definition of the last restricted simultaneous consensus object ($RSC[k]$) that the invocations $RSC[k].\mathsf{rsc\_propose}_k()$ returns at most k pairs (c_i, d_i) and those are such that $1 \leq c_i \leq k$. As for any two pairs $(c_i, prop_i)$ and $(c_j, prop_j)$ we have $(c_i = c_j) \Rightarrow (prop_i = prop_j)$, the agreement property follows. The validity and (wait-free) termination properties follow directly from the text of the algorithm and the corresponding properties of the underlying $RSC[n-1..k]$ objects. $\square_{Lemma\ 3}$

4.2 Constructing a restricted ℓ-simultaneous consensus object

Here each of an arbitrary number of processes proposes a value such that at most $\ell+1$ different values are proposed and the processes decide on at most ℓ different pairs $< c_i, d_i >$, such that $1 \leq c_i \leq \ell$, each d_i is a value that has been proposed, and any two processes that return a pair with the same c_i also return the same d_i.

The wait-free algorithm constructing a restricted ℓ-simultaneous consensus object is described in Figure 5. To reduce the number of values from $\ell+1$ to ℓ, the processes go through two sequential phases (lines 01-10, and lines 11-23). Only processes that have not decided in the first phase go into the second phase.

In the first phase (lines 01-10) the processes go through ℓ stages $T^1, ..., T^\ell$, each is one iteration of the loop in lines 02-09. A pair of arrays, $T1$ and $T2$, are associated with each stage r, $1 \leq r \leq \ell$; they are denoted $T1^r$ and $T2^r$. In each stage r, each process p_i posts its initial proposal (line 03) into $T1^r$, then takes a snapshot of the posted proposals (line 04), then posts the snapshot in $T2^r$ shared array of snapshots (line 05), and then reads all the snapshots from $T2^r$ (line 06). If a process finds a snapshot of size 1 containing some value v_j but no snapshot of size 2 then it returns the pair $< c, v_j >$, where c is the iteration number. Otherwise the process adopts the minimum value of some snapshot of size 2 or more and continues to the next iteration with this adopted value.

The key observation of the algorithm is that if a process has finished the ℓ iterations of the first phase without deciding (i.e., without returning in line 07 during any iteration), then there are snapshots of size 2 that have been posted in all the stages of the first phase. Let us notice that, due to the minimum function in line 08, one value is left behind in each iteration. Thus at most 2 different values arrive at the last (ℓ-th) iteration and, if some process did not decide in this last iteration, then the size 2 snapshot there is not empty. The size 2 snapshot in all the other iterations is also not empty because otherwise two values would have been left behind in one of the iterations, ensuring that all processes decide by the last iteration (See Lemma 5).

In the second phase (lines 11-23), all the processes that have not decided in the first phase use the vector version binary ℓ-simultaneous consensus object to decide on one of the values in these non-empty size 2 snapshots in a way that is consistent with all the decisions that have been already made during the first phase. For each stage of the first phase we associate the smaller value of the size 2 snapshot with 0, and the larger with 1. If the process also sees a snapshot of size 1 in stage r, then the r-th entry in its proposed vector is the binary value associated with the value in the size 1 snapshot (lines 14 and 15). Otherwise a process proposes an arbitrary binary value in all the other entries of its proposed binary vector (line 16). This ensures that a value that has been decided by some process during the stage r of the first phase will be the value proposed by all the processes that enter the second phase.

Finally, the binary ℓ-simultaneous consensus object is used (line 19) to decide on one of the values in these size 2 snapshots ($T2^r[2]$) and the algorithm terminates.

```
operation KSC.rsc_propose_ℓ(v_i):
(01) est_i ← v_i;
(02) for r from 1 to ℓ do
(03)     T1^r[i] ← est_i;
(04)     s_i ← snapshot_scan(T1^r);
(05)     T2^r[|s_i|] ← s_i;
(06)     for j from 1 to ℓ + 1 do ss[j] ← T2^r[j] end for;
(07)     if (ss[1] = {v} ≠ ⊥) ∧ (ss[2] = ⊥) then return(r, v)
(08)     else est_i ← min(ss[x]) for some x such that (ss[x] ≠ ⊥ ∧ x ≥ 2)
(09)     end if ;
(10) end for;
(11) for each r ∈ {1, ··· , ℓ} do
(12)   let v_m be the smallest value in T2^r[2];
(13)   let v_M be the largest value in T2^r[2];
(14)   case (T2^r[1] = {v_m}) then prop_i[r] ← 0
(15)        (T2^r[1] = {v_M}) then prop_i[r] ← 1
(16)                          else prop_i[r] ← 0 or 1 arbitrarily
(17)   end case
(18) end for;
(19) (c_i, dec_i) ← BSC[ℓ].bin_sc_propose_ℓ(prop_i); % vector version %
(20) if (dec_i = 1) then d_i ← max(T2^{c_i}[2])
(21)                else d_i ← min(T2^{c_i}[2])
(22) end if;
(23) return(c_i, d_i).
```

Figure 5: From binary ℓ-simultaneous consensus to restricted ℓ-simultaneous consensus

The rest of this section formalizes the previous intuitive presentation by proving that the algorithm described in Figure 5 implements a restricted ℓ-simultaneous consensus object.

Each cell of the shared array $T1^r$ is written at most once. It is then read through snapshot_scan() operations, and the returned snapshots are posted in $T2^r$. Thus, the sets of values associated with each snapshot form a growing sequence and each set contains all previous sets in the sequence. Hence,

Lemma 4 *For every $r, 1 \leq r \leq \ell$, for every $x \geq 1$, at most one set of values of size x is written in $T2^r[x]$ by the processes.*

The following lemma establishes that if a process does not decide in the first phase, a snapshot of size 2 has been posted in each stage r, $1 \leq r \leq \ell$ when the process starts the second phase.

Lemma 5 *In the second phase (Lines 11-23), for every $r, 1 \leq r \leq \ell$, each read of $T2^r[2]$ returns a non-$\perp$ value.*

Proof Let p_i be a process that does not decide in the first phase and starts executing the second phase. Let us assume for contradiction that the lemma is false. This means that, while p_i is executing the second phase of the protocol, a read of $T2^R[2]$ for some $R, 1 \leq R \leq \ell$ returns $\perp$. We show that p_i would have to decide in the first phase at line 07 : a contradiction.

Write, read and snapshot_scan() operations are linearizable. Let τ be the linearization point of the read of $T2^R[2]$ issued by p_i that returns $\perp$. Since no process writes $\perp$ in $T2^R[2]$, every read of $T2^R[2]$ linearized before τ must return $\perp$.

For every $r, 1 \leq r \leq \ell + 1$, let $I[r]$ be the set of values proposed to the r-th iteration in the first phase of the protocol. A value v is proposed to iteration r if v is written in some entry of $T1^r$. We claim that for every $r, 1 \leq r < \ell$, $|I[r] - I[r+1]| \geq 1$, i.e., each iteration eliminates at least one initial value (Claim *C*). Since at most $\ell + 1$ values are initially proposed, Claim *C* implies that at most $\ell + 1 - (R-1) = \ell - R + 2$ values can be written in $T1^R$. Assuming Claim *C* (which is proved in the sequel) we consider below the prefix of the execution that ends at time τ. The proof is divided in two cases according to the value of R.

- $R = \ell$. Following Claim *C* at most two values are written in $T1^\ell$, no snapshot of size ≥ 3 can be posted in $T2^\ell$. Process p_i executes iteration R before time τ. In particular, its read of $T2^\ell[2]$ returns $\perp$. It then follows from the code that the snapshot of $T1^\ell$ by p_i contains a single value, from which we conclude that p_i decides at Line 07 since it observes no posted snapshot of size ≥ 2 in $T2^\ell$.

- $R < \ell$. Each value written in $T1^{R+1}$ is the smallest value in some snapshot of size ≥ 2 that have been posted in $T2^R$. We know that at most $(\ell + 1) - (R - 1)$ values are written in $T1^R$. Therefore, no snapshot of size $\geq (\ell + 1) - (R - 1)$ is posted in $T2^R$. Moreover, before τ, no snapshot of size 2 is observed.

  Furthermore, it follows from Lemma 4 that for every $x, 2 \leq x \leq (\ell + 1) - (R - 1)$, at most one snapshot of size x can be observed in $T2^R$. Finally, since each snapshot defines a unique estimate, we conclude that at most $(\ell + 1) - (R - 1) - 2 = \ell - R$ values are proposed to iteration $R + 1$, i.e., $|I[R + 1]| \leq \ell - R$.

  It remains to show that p_i decides in the first phase. By applying Claim C to iterations $R+1, \ldots, \ell-1$, we have $|I[\ell]| \leq 1$. Process p_i executes iteration ℓ before τ. Therefore, p_i obtains a snapshot of size 1, writes it in $T2^\ell[1]$ and then decides since no snapshot of size 2 is posted in $T2^\ell$ before τ.

Claim C: $\forall r, 1 \leq r \leq \ell - 1, |I[r] - I[r+1]| \geq 1$.
Proof of Claim C. A value written in $T1^{r+1}$ is the smallest value in some snapshot of size ≥ 2 posted in $T2^r$ (Line 08). The claim follows since there are at most $|I[r]| - 1$ distinct snapshots of size ≥ 2 that may be written in $T2^r$. *End of the Proof of Claim C.* $\square_{Lemma\ 5}$

Lemma 6 *If a process decides (r, v), then $r \in \{1, \ldots, \ell\}$ and v is a value proposed by a process.*

Proof The fact that $r \in \{1, \ldots, \ell\}$ follows directly from the code of the algorithm. The validity of v follows from the observation that a value enters a snapshot only if it was already in a previous snapshot, or was proposed by a process during the first stage of the first phase. $\square_{Lemma\ 6}$

Lemma 7 *If p_i and p_j decide (r_i, v_i) and (r_j, v_j), respectively, we have $(r_i = r_j) \Rightarrow (v_i = v_j)$.*

Proof For each consensus instance R, let D_R denote the set of processes that decide in the r-th consensus instance. We consider three cases according to the phase(s) in which processes that belong to D_R decide.

- All the processes that belong to D_R decide in the first phase (Line 07). In that case, process $p_i \in D_R$ decides a value contained in a singleton snapshot that it has observed in $T2^R$. Agreement follows from the fact that a unique snapshot of size one may be posted in $T2^R$ by the different processes (Lemma 4).

- All the processes that belong to D_R decide in the second phase (line 23). Each process $p_i \in D_R$ gets back a pair (R, d_i) from the binary ℓ-simultaneous consensus object. Due to the agreement property of the object, $\exists d \in \{0, 1\}$ such that $\forall p_i \in D_R, d_i = d$.

  Moreover, per Lemma 5, every process in D_R observes a snapshot of size 2 in $T2^r$ at lines 20 or 21 and, by Lemma 4, they observe the same snapshot. It then follows from lines20-22 that each process in D_R returns the same value.

- Decisions occur in both phases. Let C be the set of processes that invoke the binary ℓ-simultaneous consensus object (a process that belongs to C could have not decided in the first phase). Among them, let p_c be the first process that reads $T2^R[1]$ in the second phase of the algorithm (lines 14-15). This occurs at time τ. There are two cases according to the value returned by that read.

  - Suppose that p_c does not observe a snapshot of size 1 ($T2^R[1] = \perp$). In that case no process in D_R could decide in the first phase of the algorithm.

    Assume for contradiction that process p_i decides (R, v) at line 07. p_i must observe $T2^R[1] = \{v\}$ at some time τ'. As the read of $T2^R[1]$ by p_c returns $\perp$, and no process writes $\perp$ in $T2^R[1]$, it follows that $\tau' > \tau$. But by Lemma 5, we know that $T2^R[2] \neq \perp$ when p_c starts the second part of the protocol. Consequently, p_i must also observe $T2^R[2] \neq \perp$, which prevents it from deciding in the first part of the protocol (line 07).

  - Suppose that p_c observes a singleton snapshot $\{v\}$ in $T2^R$. Per Lemma 4, only one singleton snapshot can be written in $T2^R$. Therefore, every process in C reads $\{v\}$ in $T2^R[1]$ at lines 14-15. Then every process in C "proposes" v to the Rth binary-consensus. More precisely, each process proposes $d = 0$ (resp. $d = 1$) to the Rth binary consensus if v is the smallest (resp. greatest) value in the snapshot written in $T2^R[2]$ (by Lemma 5, there is always a snapshot s of size 2 written in $T2^R[2]$ when processes execute the second part of the protocol. Since snapshots are ordered by containment, $v \in s$).

    Therefore, by the validity property of the ℓ-simultaneous binary consensus object, each process in D_R that decides in the second part gets back (R, d) from the object, and consequently returns the same pair (R, v) (lines 20-22). Moreover, a process in D_R that decides in the first part of the protocol returns also (R, v), $\{v\}$ being the only snapshot written in $T2^R[1]$.

$\square_{Lemma\ 7}$

Lemma 8 *The algorithm described in Figure 5 is a wait-free construction of a restricted ℓ-simultaneous consensus object from a binary vector ℓ-simultaneous consensus object for any number of processes.*

Proof The wait-free property follows directly from the text of the algorithm and the same property of the underlying binary simultaneous consensus object. The validity and the agreement properties have been proved in Lemma 6 and Lemma 7, respectively. $\square_{Lemma\ 8}$

4.3 A second equivalence

Theorem 2 *The multivalued k-simultaneous consensus problem and the binary k-simultaneous consensus problem are wait-free equivalent in read/write shared memory systems made up of an arbitrary number of processes.*

Proof As already indicated, the multivalued version of the problem trivially solves its binary version. The other direction follows from the algorithm described in Figure 4 (proved in Lemma 3), and the algorithm described in Figure 5 (proved in Lemma 8). $\square_{Theorem\ 2}$

5 Conclusion

This paper has introduced and studied the k-simultaneous consensus problem. Its main result is the following theorem, whose proof follows from Theorem 1 and Theorem 2.

Theorem 3 *The k-set agreement problem and the k-simultaneous binary consensus problem are wait-free equivalent in asynchronous read/write shared memory systems made up of an arbitrary number of processes.*

This theorem not only provides a new characterization of the k-set agreement problem, but also gives a precise meaning to the notion of binary k-set agreement. In that sense, it shows in a clear way that k-simultaneous consensus captures both k-set agreement and consensus. Hence, it is a stronger and more general paradigm than k-set agreement.

A natural generalization of ℓ-simultaneous consensus is the ℓ-simultaneous k-set-agreement [2, 13]. This problems is defined in the same way as the ℓ-simultaneous consensus problem, namely, each process has to decide a pair (c, v) subject to the following constraints : (1) $1 \leq c \leq \ell$, (2) v is a proposed value for the c-th instance and (3) at most k values are decided in each instance. It is easy to see that the scalar version an vector version of this problem are equivalent. Also, given a solution to the ℓ-simultaneous k-set-agreement problem, it easy to solve $k\ell$-set-agreement, since at most $k\ell$ pairs are decided.

What about the other direction ? A simple modification of the algorithm described in Figure 3 constructs a ℓ-simultaneous k-set-agreement object from a $k\ell$-set-agreement object. In a very simple way, the first statement in line 04 that defines the consensus instance is replaced by "**let** $c_i = \lceil \frac{|snap_i|}{k} \rceil$" that now defines the k-set instance number associated with the value decided by p_i.

The $k\ell$-set-agreement object reduces the number of distinct values to $k\ell$. Thus, the first coordinate of the decided pairs is at most ℓ. Finally, as snapshots are related by containment, there at most k distinct snapshots $snap$ such that $(c-1)k+1 \leq |snap| \leq ck$. Therefore, at most k values are decided in the cth instance. Hence,

Theorem 4 *The $k\ell$-set agreement problem and the ℓ-simultaneous k-set-agreement problem are wait-free equivalent in asynchronous read/write shared memory systems made up of an arbitrary number of processes.*

References

[1] Afek Y., Attiya H., Dolev D., Gafni E., Merritt M., and Shavit N., Atomic Snapshots of Shared Memory. *Journal of the ACM*, 40(4):873-890, 1993.

[2] Afek Y., Gafni E., Rajsbaum S., Raynal M. and Travers C., Simultaneous Consensus Tasks: a Tighter Characterization of Set Consensus. *Proc. 8th Int'l Conference on Distributed Computing and Networking (ICDCN'06)*, Springer-Verlag LNCS #4308, pp. 331-341, 2006.

[3] Aguilera M., A pleasant stroll through the land of infinitely many creatures. *ACM SIGACT News, Distributed Computing Column*, 35(2):36-59, 2004.

[4] Attiya H. and Welch J., *Distributed Computing: Fundamentals, Simulations and Advanced Topics*, (2d Edition), Wiley-Interscience, 414 pages, 2004.

[5] Borowsky E. and Gafni E., Generalized FLP Impossibility Results for t-Resilient Asynchronous Computations. *Proc. 25th ACM Symposium on Theory of Computing (STOC'93)*, pp. 91-100, 1993.

[6] Borowsky E., Gafni E., Lunch N. an Rajsbaum S., The BG Distributed Simulation Algorithm. *Distributed Computing*, 14:127-146, 2001.

[7] Chaudhuri S., More *Choices* Allow More *Faults:* Set Consensus Problems in Totally Asynchronous Systems. *Information and Computation*, 105:132-158, 1993.

[8] Dolev D., Reischuk R. and Strong R., Early Stopping in Byzantine Agreement. *Journal of the ACM*, 37(4):720-741, April 1990.

[9] Dwork C. and Moses Y., Knowledge and Common Knowledge in a Byzantine Environment: Crash Failures. *Information and Computation*, 88(2):156-186, 1990.

[10] Fischer M.J., Lynch N.A. and Paterson M.S., Impossibility of Distributed Consensus with One Faulty Process. *Journal of the ACM*, 32(2):374-382, 1985.

[11] Gafni E., Merritt M., and Taubenfeld G., The Concurrency Hierarchy, and Algorithms for Unbounded Concurrency. *Proc. 20th ACM Symposium on Principles of Distributed Computing (PODC'01)*, ACM Press, pp. 161-170, 2001.

[12] Gafni E. and Rajsbaum S., Musical Benches. *Proc. 19th Int'l Symposium on Distributed Computing (DISC'05)*, Springer Verlag LNCS #3724, pp. 63–77, 2005.

[13] Gafni E., Rajsbaum R., Raynal M. and Travers C., The Committee Decision Problem. *Proc. 8th Latin American Theoretical Informatics (LATIN'06)*, Springer-Verlag LNCS #3887, pp. 502-514, 2006.

[14] Herlihy M.P., Wait-Free Synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124-149, 1991.

[15] Herlihy M.P. and Shavit N., The Topological Structure of Asynchronous Computability. *Journal of the ACM*, 46(6):858-923, 1999.

[16] Herlihy M.P. and Wing J.M., Linearizability: a Correctness Condition for Concurrent Objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463-492, 1990.

[17] Lamport L., On interprocess communication, Part 1: Models, Part 2: Algorithms. *Distributed Computing*, 1(2):77-101, 1986.

[18] Loui M.C., Abu-Amara H., Memory Requirements for Agreement Among Unreliable Asynchronous Processes. *Advances in Computing research*, JAI Press, 4:163-183, 1987.

[19] Lynch N.A., Distributed Algorithms. *Morgan Kaufmann Pub.*, San Francisco (CA), 872 pages, 1996.

[20] Merritt M. and Taubenfeld G., Computing with Infinitely Many Processes. *Proc. 14th Int'l Symposium on Distributed Computing (DISC'00)*, Springer-Verlag LNCS #1914, pp. 164-178, 2000.

[21] Mostéfaoui A., Raynal M. and Tronel F., From Binary Consensus to Multivalued Consensus in Asynchronous Message-Passing Systems. *Information Processing Letters*, 73:207-213, 2000.

[22] Saks M. and Zaharoglou F., Wait-Free k-Set Agreement is Impossible: The Topology of Public Knowledge. *SIAM Journal on Computing*, 29(5):1449-1483, 2000.

[23] Zielińsky P., Anti-Ω: the Weakest Failure Detector for Set Agreement. *Proc. 27th ACM Symposium on Principles of Distributed Computing (PODC'08)*, ACM Press, pp. 55-64, 2008.