Consensus in Byzantine Asynchronous Systems

Roberto Baldoni, Jean-Michel Hélary, Michel Raynal, Lénaïck Tanguy

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Consensus in Byzantine Asynchronous Systems

Roberto BALDONI , Jean-Michel HÉLARY , Michel RAYNAL , Lénaïck TANGUY

N ° 3655

THÈME 1

ISSN 0249-6399

Consensus in Byzantine Asynchronous Systems

Roberto BALDONI* , Jean-Michel HÉLARY† , Michel RAYNAL‡ , Lénaïck TANGUY§

Thème 1 — Réseaux et systèmes
Projet ADP

Rapport de recherche n3655 — — 21 pages

Abstract: This paper studies the consensus problem in byzantine asynchronous distributed systems. In such systems, a process may stop communicating with other processes or may behave arbitrarily (e.g., execute a statement more than once, corrupt the value of a local variable or misevaluate a local predicate).

A consensus protocol resilient to such failures is proposed. It uses signed and certified messages and is based on two underlying failure detection modules. The first is an unreliable failure detector module . The second is a reliable arbitrary behavior detection module. More precisely, the first module detects processes that stop sending messages, while processes experiencing other arbitrary behaviors are detected by the second module. The protocol is resilient to F faulty processes, where F is less than half of the total number of processes and less than an upper bound C (where C is the maximum number of faulty processes that can be tolerated by the underlying certification service).

The approach used to design the protocol is new. While usual byzantine consensus protocols are based on unreliable failure detectors to detect processes that stop communicating, none of them use a module to detect their arbitrary behavior (this detection is not isolated from the protocol and makes it difficult to understand and prove correct). In addition to this modular approach and to a new consensus protocol for byzantine systems, the paper presents a finite state automaton-based implementation of the arbitrary behavior detection module. Finally, the modular approach followed in this paper can be used to solve other problems in byzantine systems.

Key-words: Asynchronous Systems, Consensus Problem, Byzantine Failures, Fault-Tolerance, Unreliable Failure Detectors.

(Résumé : tsvp)

* DIS, Universitá di Roma *La Sapienza*, Via Salaria 113, Roma, Italy. baldoni@dis.uniroma1.it
† IRISA. helary@irisa.fr
‡ IRISA. raynal@irisa.fr
§ IRISA. ltanguy@irisa.fr

Unit´e de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
T´el´ephone : 02 99 84 71 00 - International : +33 2 99 84 71 00
T´el´ecopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Consensus dans les systèmes asynchrones byzantins

Résumé : Dans ce papier on s'intéresse au problème du consensus dans un système réparti asynchrone soumis à des défaillances byzantines. Dans de tels systèmes, un processus peut arrêter de communiquer avec d'autres processus ou peut se comporter de manière arbitraire (par exemple exécuter plusieurs fois une instruction, corrompre la valeur d'une variable locale, évaluer incorrectement une condition, etc).

On propose un protocole de consensus tolérant de telles défaillances. D'une part, il utilise des messages signés et certifiés. D'autre part, il s'appuie sur deux détecteurs de défaillances. Le premier est un détecteur non fiable, qui détecte les processus suspectés d'arrêter d'envoyer des messages. Le second est un détecteur fiable qui détecte les autres comportements arbitraires. Le protocole tolère un nombre maximum de F processus incorrects, où F est inférieur à la moitié du nombre de processus et à une borne C imposée par le service de certification.

Cette approche est nouvelle. Alors que d'autres protocoles de consensus s'appuient sur des détecteurs non fiables pour suspecter les processus ayant arrêté de communiquer, aucun n'utilise de module de détection de comportement arbitraire (cette détection est incluse dans les protocoles, ce qui les rend plus difficile à comprendre et à prouver). En plus de cette approche modulaire et de la présentation d'un nouveau protocole de consensus en environnement byzantin, on présente une implémentation du détecteur de comportements arbitraires basée sur des automates. Enfin, l'approche méthodologique proposée dans ce papier peut être appliquée à la résolution d'autres problèmes en environnement byzantin.

Mots-clé : Systèmes asynchrones, consensus, comportement byzantin, tolérance aux fautes, détecteurs de défaillances non fiables.

1 Introduction

Consensus is a fundamental paradigm for fault-tolerant asynchronous distributed systems. Each process proposes a value to the others. All correct processes have to agree (Termination) on the same value (Agreement) which must be one of the initially proposed values (Validity).

Solving consensus in asynchronous distributed systems where processes can crash is a well-known difficult task: Fischer, Lynch and Paterson have proved an impossibility result [5] stating that *there is no deterministic solution to the consensus problem in the presence of even a single crash failure*. A way to circumvent this impossibility result is to use the concept of unreliable failure detectors introduced by Chandra and Toueg [2]. Each process is equipped with a failure detector module that provides it with a list of processes currently suspected by the detector to have crashed. A failure detector module can make mistakes by not suspecting a crashed process or by erroneously suspecting a correct one. Formally, a failure detector module is defined by two properties: completeness (a property on the actual detection of process crashes), and accuracy (a property that restricts the mistakes on erroneous suspicions). Among those classes, the weakest one that allows to solve the consensus problem in a process crash failure model, denoted $\diamond\mathcal{S}$, is based on strong completeness (eventually every crashed process is detected by every correct process) and eventual weak accuracy (there is a time after which some correct process is not suspected by any correct process).

Solving consensus in an environment where processes can exhibit arbitrary behavior (e.g., omit to execute a statement or corrupt the value of a local variable) is notably difficult. First steps in this direction have been carried out by Malkhi and Reiter [9], Kihlstrom, Moser and Melliar-Smith [8] and Doudou and Schiper [4].

Malkhi and Reiter [9] have been the first to address this problem, by extending failure detectors to cope with byzantine failures. They have introduced a failure detector class $\diamond\mathcal{S}(bz)$ based on the notion of *quiet* process. A process p is quiet if there does exist a time after which some correct process does not receive anymore messages from p. (Note that a crashed process is a quiet process). $\diamond\mathcal{S}(bz)$ is a failure detector with a strong completeness property and an eventual weak accuracy property. The latter property is similar to the one given in the crash failure model (note that this property does not concern non-correct processes). The strong completeness property must be reformulated to include the notion of *non-correct process* which is dependent on the considered failure model. Here, strong completeness means that each correct process eventually suspects each quiet process.

Kihlstrom, Moser and Melliar-Smith [8] have pointed out that messages need to be *certified* as well as *signed*. The signature allows a receiver to verify the sender while the certificate is a well-defined amount of redundant information carried by messages that allows a receiver to check if the content of a message is valid and if the sending of the message was done "at the right time". In other word, a certificate allows a receiver to "look into the sender process" in order to see if the actions that produced the sending were correct.

Both previous solutions [8, 9] solve Strong Consensus, defined by the traditional agreement and termination properties, but with another validity property, named Strong Validity. It stipulates that, *if all correct processes propose the same input value v, then a correct process that decides must decide on that value*. Strong Consensus has a major drawback: when correct processes propose different values, they are allowed to decide on a value that is not necessarily connected to their inputs. More generally, the traditional validity property is not adequate to define the consensus problem in a byzantine setting [4]. A byzantine process can initially propose an irrelevant value (i.e., a value different from the one it should propose) and then behave correctly. There is no way for the other processes to detect this failure. Consequently, the set of correct processes could agree on an irrelevant value v proposed by a process. A significant advance has been done by Doudou and Schiper to circumvent this drawback: they have introduced a new validity property, namely the *Vector Validity* property. In this case, each process proposes a vector which contains a certain number (at least one) of correct entries (this problem is called *vector consensus*). An entry is correct if it is from a correct process. So, processes have first to construct these vectors. Finally, processes agree on one of these vectors. They have also shown that Vector Consensus does not suffer from the same drawback as Strong Consensus, in the sense that other agreement problems (e.g., Atomic Broadcast) can be reduced to Vector Consensus. They have also defined a failure detector class ($\diamond\mathcal{M}$) based on the notion of *Mute* process, close to the notion of *quiet* process introduced by Malkhi and Reiter. Using $\diamond\mathcal{M}$ and properly signed and certified messages, they have proposed a protocol that solves Vector Consensus. Let F be the maximal number of processes that may be faulty (with ($F < n$ where n is the total number of processes). Doudou and Schiper's protocol satisfies Agreement, Termination and Vector Validity where at least $F + 1$ entries of the decided vector are from correct processes. The failure detector $\diamond\mathcal{M}$ is unreliable, and detects only mute processes. The detection of other byzantine failures (sta-

tement omission, value corruption, etc.) is not performed by this failure detector, and is left to the protocol itself.

All the previous protocols (Malkhi and Reiter [2], Kihlstrom, Moser and Melliar-Smith [8] and Doudou and Schiper [4]) require $F = \lfloor (n-1)/3 \rfloor$ ([1]). We propose in this paper a protocol to solve Vector Consensus in a byzantine asynchronous distributed system. This protocol is based on a failure detector of the class $\diamond\mathcal{S}(bz)$, and uses signed and certified messages. It assumes $F \leq \min(\lfloor (n-1)/2 \rfloor, C)$ where C is the maximum number of faulty processes the underlying certification service used by the protocol can cope with[2]. Our approach clearly separates the constraints due to the underlying certification mechanism from the constraints due to the consensus protocol that uses this mechanism. The protocol uses as a skeleton a consensus protocol designed by Hurfin and Raynal [6] for the crash failure model, and extends it to cope with the following "bad" process behaviors: crash, omission of a statement, multiple execution of a statement, corruption of a local variable and misevaluation of a local predicate. The resulting protocol satisfies Agreement, Termination and Vector Validity with at least $\alpha = n - 2F$ entries from correct processes (note that, due to definition of F, we have $\alpha \geq 1$). So, the proposed protocol allows to solve other agreement problems, such as Atomic Broadcast.

In the proposed protocol, each process is actually composed of five modules: (i) a consensus module, (ii) an unreliable failure detection module of the class $\diamond\mathcal{S}(bz)$, (iii) an arbitrary behavior detection module, (iv) a certification module and (v) a signature module. The consensus module executes the protocol. The unreliable failure detection module of the class $\diamond\mathcal{S}(bz)$ and the arbitrary behavior detection module are used to reveal quiet processes, and processes with other byzantine behaviors, respectively. If the process is byzantine, these three modules can behave in a byzantine way. The other two modules do not behave maliciously. The signature module filters out messages in which the sender must be identified. The certification module manages certificates to be associated with messages. Note that the previous byzantine consensus protocols leave to the protocol itself the detection of faulty processes not captured by the unreliable failure detector $\diamond\mathcal{S}(bz)$ (or $\diamond\mathcal{M}$). The introduction of an arbitrary behavior detection module, separated from the consensus protocol, is a new approach that provides a general, efficient, modular and simple way to extend distributed protocols to cope with byzantine failures. In that sense, the contribution of this paper is not only algorithmic and practical (with the design of a new byzantine consensus protocol), but also methodological.

Moreover, an implementation of the arbitrary behavior detection module is described. This implementation is based on a set of finite state automata, one for each process. At the operational level, the module associated with a process p_i intercepts all messages sent to p_i from the underlying network and checks if their sendings are done according to the program specification. In the affirmative, it relays the message to p_i's consensus module. This module maintains a set ($byzantine_i$) of processes it detected to experience at least one byzantine behavior. Differently from the failure detector module $\diamond\mathcal{S}(bz)$, the arbitrary behavior detection module is reliable (i.e., if the process to which it belongs is correct, it does not make mistake).

The paper is made of seven sections. Section 2 addresses the consensus problem in a crash failure model. Section 3 presents the byzantine asynchronous distributed system model. Then Section 4 presents the byzantine consensus protocol. Section 5 provides a finite state automaton-based implementation of the arbitrary behavior detection module. Section 6 proves the correctness of the proposed consensus protocol. Finally, Section 7 concludes the paper.

2 Consensus in the Crash Failure Model

This section considers the consensus problem in a distributed asynchronous system where processes can fail by crashing (i.e., a process behaves correctly until it possibly crashes). So, in this section, a *correct* process is a process that does not crash. Failure detectors suited to this model [2] and a consensus protocol [6] based on such failure detectors are presented.

[1] In an asynchronous distributed system prone to process crash failures and equipped with a failure detector of the class $\diamond\mathcal{W}$, the *consensus* problem requires $F \leq \lfloor (n-1)/2 \rfloor$ to be solved [2]. In synchronous systems with byzantine process failures, the *byzantine general* problem requires $F \leq \lfloor (n-1)/3 \rfloor$ if messages are not signed ("oral messages"), and $F \leq (n-1)$ if messages are signed [10].

[2] Usual certification mechanisms require $C = \lfloor (n-1)/3 \rfloor$. This explains why previous works consider $F = \lfloor (n-1)/3 \rfloor$ in their consensus protocols and in their proofs.

2.1 Asynchronous Systems

We consider a system consisting of $n > 1$ processes $\Pi = \{p_1, p_2, \ldots, p_n\}$. Each process executes a sequence of statements defined by its program text. Processes communicate by exchanging messages through *reliable*[3] and FIFO[4] channels. As the system is asynchronous, there is no assumption about the relative speed of processes or the message transfer delays.

2.2 Consensus

Every correct process p_i *proposes* a value v_i and all correct processes have to *decide* on some value v, in relation to the set of proposed values. More precisely, the *Consensus problem* is defined by the three following properties [2, 5]:

- Termination: Every correct process eventually decides on some value.
- Validity: If a process decides v, then v was the initial value proposed by some process.
- Agreement: No two correct processes decide differently.

2.3 Failure Detectors

Informally, a failure detector consists of a set of modules, each one attached to a process: the module attached to p_i maintains a set (named $suspected_i$) of processes it currently suspects to have crashed. Any failure detector module is inherently unreliable: it can make mistakes by not suspecting a crashed process or by erroneously suspecting a correct one. As in [2], we say "process p_i suspects process p_j" at some time t, if at time t we have $p_j \in suspected_i$.

Failure detector classes have been defined by Chandra and Toueg [2] in terms of two properties, namely *Completeness* and *Accuracy*. They showed that the weakest failure detectors to solve consensus is $\diamond\mathcal{S}$ which is based on the following properties:

- Strong Completeness: Eventually, every crashed process is permanently suspected by every correct process.
- Eventual Weak Accuracy: There is a time after which some correct process is never suspected by any correct process.

Since then many protocols solving consensus have been proposed using failure detectors $\diamond\mathcal{S}$. These protocols, in the setting proposed in this section, require the majority of processes to be correct (i.e., $F < n/2$, where F is the maximum number of faulty processes).

2.4 Hurfin-Raynal's Consensus Protocol

As indicated in the introduction, several protocols solving consensus in byzantine systems have been proposed. Actually each of them is based on a skeleton protocol solving consensus in a process crash model. The protocol proposed by Malkhi and Reiter [2] and the one proposed by Kihlstrom, Moser and Melliar-Smith [8] use [2] as a skeleton. The protocol proposed by Doudou and Schiper [4] uses [11] as a skeleton. We have chosen [6] as a skeleton for our byzantine consensus protocol because, in addition to its conceptual simplicity, this protocol is particularly efficient when the underlying failure detector makes no mistakes, whether there are failures or not (see [6] for more details). This rest of this section provides a brief description of a version of this protocol that assumes FIFO channels (this constraint is not required by the original protocol).

As other consensus protocols, this one proceeds in successive asynchronous rounds and uses the rotating coordinator paradigm. During a round, a predetermined process (the round coordinator) tries to impose a value as the decision value. To attain this goal, each process votes: either (vote CURRENT) in favor of the value proposed by the round coordinator (when it has received one), or (vote NEXT) to proceed to the next round and benefit from a new coordinator (when it suspects the current coordinator).

[3] i.e., a message sent by a process p_i to a process p_j is eventually received by p_j, if p_j is correct.
[4] This assumption simplifies the solution when addressing byzantine failures.

Automaton States During each round, the behavior of each process p_i is determined by a finite state automaton. This automaton is composed of 3 states. The local variable $state_i$ will denote the automaton state in which p_i currently is. During a round, the states of the automaton have the following meaning:

- $state_i = q_0$: p_i has not yet voted (q_0 is the automaton initial state).
- $state_i = q_1$: p_i has voted CURRENT and has not changed its mind (p_i moves from q_0 to q_1).
- $state_i = q_2$: p_i has voted NEXT.

Local variables In addition to the local variable $state_i$, process p_i manages the following four local variables:

- r_i defines the current round number.
- est_i contains the current estimation by p_i of the decision value.
- $nb\_current_i$ (resp. nb_next_i) counts the number of CURRENT (resp. NEXT) votes received by p_i during the current round.
- rec_from_i is a set composed of the process identities from which p_i has received a (CURRENT or NEXT) vote during the current round.

Finally, $suspected_i$ is a set managed by the associated failure detector module (cf. Section 2.3); p_i can only read this set.

Automaton Transitions The protocol manages the progression of each process p_i within its automaton, according to the following rules. At the beginning of round r, $state_i = q_0$. Then, during r, the transitions are:

- *Transition* $q_0 \rightarrow q_1$ (p_i *first votes CURRENT*). This transition occurs when p_i, while in the initial state q_0, receives a CURRENT vote (line 8). This means that p_i has not previously suspected the round coordinator (line 13). Moreover, when p_i moves to q_1 and it is not the current coordinator, it broadcasts a CURRENT vote (line 11).
- *Transition* $q_0 \rightarrow q_2$ (p_i *first votes NEXT*). This transition occurs when p_i, while in the initial state q_0, suspects the current coordinator (line 13). This means that p_i has not previously received a CURRENT vote. Moreover, when p_i moves to q_2, it broadcasts a NEXT vote (line 14).
- *Transition* $q_1 \rightarrow q_2$ (p_i *changes its mind*). This transition (executed by statements at line 18) is used to prevent a possible deadlock. A process p_i that has issued a CURRENT vote is allowed to change its mind if p_i has received a (CURRENT or NEXT) vote from a majority of processes (*i.e.*, $\mid rec\_from_i \mid > n/2$) but has received neither a majority of CURRENT votes (so it cannot decide), nor a majority of NEXT votes (so it cannot progress to the next round). Then p_i changes its mind in order to make the protocol progress: it broadcasts a NEXT vote to favor the transition to the next round (line 18).

Protocol Description Function consensus() consists of two concurrent tasks. The first task handles the receipt of a DECIDE message (lines 2-3); it ensures that if a process p_i decides (line 3 or line 12), then all correct processes will also receive a DECIDE message. The second task (lines 4-21) describes a round: it consists of a loop that constitutes the core of the protocol. Each (CURRENT or NEXT) vote is labeled with its round number[5].

- At the beginning of a round r, the current coordinator p_c proposes its estimate v_c to become the decision value by broadcasting a CURRENT vote carrying this value (line 6).
- Each time a process p_i receives a (CURRENT or NEXT) vote, it updates the corresponding counter and the set rec_from_i (lines 9 and 16).
- When a process receives a CURRENT vote for the first time, namely, CURRENT(p_k,r,est_k), it adopts est_k as its current estimate est_i (line 11). If, in addition, it is in state q_0, it moves to state q_1.

[5]In any round r_i, only votes related to round r_i can be received. A vote from p_k related to a past round is discarded and a vote related to a future round r_k (with $r_k > r_i$) is buffered and delivered when $r_i = r_k$.

```
function consensus(v_i)

(1) r_i ← 0; est_i ← v_i;
    cobegin
(2)     || upon receipt of DECIDE(p_k, est_k)
(3)        send DECIDE(p_i, est_k) to Π; return(est_k)

(4)     || loop % on a sequence of asynchronous rounds %
(5)        c ← (r_i mod n) + 1; r_i ← r_i + 1; state_i ← q_0; rec_from_i ← ∅; nb_next_i ← 0; nb_current_i ← 0;
(6)        if (i = c) then send CURRENT(p_i, r_i, est_i) to Π; state_i ← q_1 endif;

(7)        while (nb_next_i ≤ n/2) do % wait until a branch can be selected, and then execute it %
(8)           upon receipt of CURRENT(p_k, r_i, est_k)
(9)                 nb_current_i ← nb_current_i + 1; rec_from_i ← rec_from_i ∪ {p_k};
(10)                if (nb_current_i = 1) then est_i ← est_k endif;
(11)                if (state_i = q_0) then send CURRENT(p_i, r_i, est_i) to Π; state_i ← q_1 endif;
(12)                if (nb_current_i > n/2) then send DECIDE(p_i, est_i) to Π; return(est_i) endif

(13)          upon (p_c ∈ suspected_i)
(14)                if (state_i = q_0) then state_i ← q_2; send NEXT(p_i, r_i) to Π endif

(15)          upon receipt of NEXT(p_k, r_i)
(16)                nb_next_i ← nb_next_i + 1; rec_from_i ← rec_from_i ∪ {p_k}

(17)          upon ((state_i = q_1) ∧ (| rec_from_i |> n/2))
(18)                state_i ← q_2; send NEXT(p_i, r_i) to Π
(19)       endwhile

(20)       if (state_i ≠ q_2) then state_i ← q_2; send NEXT(p_i, r_i) to Π endif
(21)    endloop
    coend
```

Figure 1: Hurfin-Raynal's $\diamond\mathcal{S}$-based Consensus Protocol (adapted to FIFO channels)

- A process p_i decides on an estimate proposed by the current coordinator as soon as it has received a majority of CURRENT votes, *i.e.*, a majority of votes that agree to conclude during the current round (line 12).
- When a process progresses from round r to round $r+1$ it issues a NEXT (line 20) if it did not do it in the **while** loop. These NEXT votes are used to prevent other processes from remaining blocked in round r (line 7).

3 Consensus in a Byzantine Model

This section first defines what is meant by "byzantine" behavior. Then the section defines a version of the consensus problem suited to the byzantine system model. Finally the modules that are part of the system model are described.

3.1 Byzantine Processes

The previous section has assumed processes fail only by crashing. Here, we define the model of a byzantine process. This is actually a quite important step in order to design, in a systematic way, mechanisms allowing a distributed protocol to safely progress in the presence of such failures. In this section, a *correct* process is a process that does not exhibit a byzantine behavior. A process is byzantine if, during its execution, one of the following faults occurs:

Crash The process stops executing statements of its program and halts.

Corruption The process changes arbitrarily the value of a local variable (i.e., arbitrary assignment) with respect to its program specification. This fault could be propagated to other processes by including incorrect values in the content of a message sent by the process.

Omission The process omits to execute a statement of its program (e.g., it omits to send a message, it omits to update a variable, etc.). If a process omits to execute an assignment, this could lead to a corruption fault.

Duplication The process executes more than one time a statement of its program. Note that if a process executes an assignment more than one time, this could lead to a corruption fault.

Misevaluation The process misevaluates an expression included in its program. This fault is different from a corruption fault: misevaluating an expression does not imply the update of the variables involved in the expression and, in some cases (e.g., conditions used in **if** and **loop** statements) the result of an evaluation is not assigned to a variable.

Previous distinction between misevaluation and corruption is extremely important, particularly in the case of conditions. In fact conditions involving local variables can be misleading because of a corruption fault, even though no misevaluation occurred. For example, in a test like **if** $(state_i = q_0)$ (line 14 of Figure 1), the value $state_i$ could have been previously corrupted. In other words the comparison might occur in a corrupted context. If conditions would involve only values that cannot be corrupted (e.g., values of constant) then only a misevaluation fault could lead to a wrong result.

3.2 From Consensus to Vector Consensus

Doudou and Schiper have pointed out [4] that in the presence of byzantine processes the validity property of consensus described in Section 2.2 is not adequate (see the discussion in the introduction). Hence, they have defined the *Vector Consensus* problem[6] by using the Agreement and Termination property of Section 2.2 and the following Vector Validity property [4]:

Vector Validity Every process decides on a vector $vect$ of size n:

- For every process p_i: if p_i is correct, then either $vect[i] = v_i$ or $vect[i] = null$, and
- At least $\alpha \geq 1$ elements of $vect$ are initial values of correct processes[7].

3.3 The Class $\diamond\mathcal{S}(bz)$ of Failure Detectors

The Eventual Weak Accuracy property can also be used in a byzantine setting as it is based on the notion of correct processes. The Completeness property needs to be re-stated as it is based on the notion of non-correct processes which, in a byzantine environment, has a wider meaning.

Let us consider the concept of quiet process introduced by Malkhi and Reiter [9]. Considering a quiet process as a faulty process, the Completeness property of Section 2.3 is redefined as follows [9]:

- Strong Completeness: Eventually, every quiet process is permanently suspected by every correct process.

A failure detector that satisfies Eventual Weak Accuracy and the definition of Strong Completeness based on the notion of quiet process belongs to the $\diamond\mathcal{S}(bz)$ class. It is interesting to remark that the *quietness* notion captures only some byzantine behaviors that lead the protocol to omit all sendings of messages. All the other byzantine behaviors are captured by the arbitrary behavior detection module that will be introduced in the next section.

3.4 Additional Assumptions on the Model

We denote by F the maximum number of byzantine processes. F is a known bound, and we assume $F \leq \min(\lfloor \frac{n-1}{2} \rfloor, C)$ (where C is the maximum number of faulty processes allowed by the certification mechanism). So, at least $n - F \geq \max(\lceil \frac{n+1}{2} \rceil, n - C)$ processes are correct.

[6] The *Vector Consensus* notion has first been proposed in synchronous systems where it is called *Interactive Consistency* problem [10]. This notion has also been investigated in asynchronous systems with process crash failures in [7].

[7] [4] considers the case $\alpha = F + 1$.

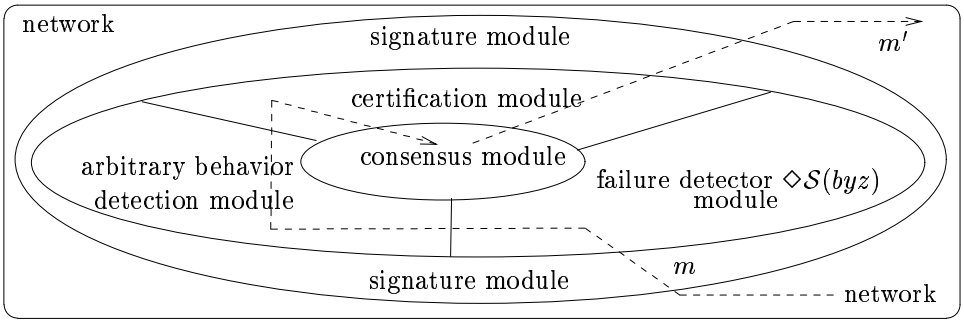

Figure 2: Structure of a process p_i.

Key Cryptosystem Each process p_i possesses a private key and a public key. The private key is used by p_i to sign, in an unforgeable way, outgoing messages. A message m signed by p_i is denoted $\langle m \rangle_i$. Upon the arrival of a signed message at p_i, the sender's public key allows the receiver p_i to verify the identity of the assumed message sender.

Certificates The usage of a key cryptosystem does not allow a receiver (i) to verify the correctness of the content of a message nor (ii) to check if the sender process has correctly followed its program specification.

To address these two problems, each message contains a *certificate* which represents a properly formed set of messages. This set is used by the receiver to check the correctness of the content of a message and the proper sending of the message. Hence, with each type of messages used by the consensus protocol we have to associate a proper certificate that will allow a receiver to check points (i) and (ii). To check point (i), usual certification techniques are based on "majority" rules. So, they require that proper certificates contain at least $n - C$ messages, able to "witness" the content of the message. Usually, $n - C = \lceil \frac{2n+1}{3} \rceil$.

We denote as $\langle m, cert_m \rangle_i$ a message m signed and sent, with the certificate $cert_m$, by p_i. $\langle m, cert_m \rangle_i$ is *properly formed* if its certificate is a properly formed set of messages. Each certificate is updated in a safe way (i.e., no corruption fault can occur) by a process (i.e., if a process forges a certificate, it will be immediately detected by the receiver as a byzantine process) and a process can safely query about the cardinality of a given certificate. Note that the fact that p_i cannot corrupt a given certificate does not prevent p_i's misevaluation faults.

3.5 Structure of a Process

A process consists of five modules whose role has been presented in the introduction: (i) a consensus module, (ii) an arbitrary behavior detection module, (iii) an unreliable failure detection module of the class $\diamond\mathcal{S}(bz)$, (iv) a certification module, and (v) a signature module. More precisely, the structure of a process p_i is given in Figure 2. The same figure also shows the path followed by a message m (resp. m') received (resp. sent) by p_i.

Signature module Each message arriving at p_i is first processed by this module which verifies the signature of the sender (by using its public key). If the signature of the message is inconsistent with the identity field contained in the message, the message is discarded and its sender identity (known thanks to the unforgeable signature), is passed to the arbitrary behavior detection module to be added to the set $byzantine_i$. Otherwise, the message is passed to the local failure detection module $\diamond\mathcal{S}(bz)$. Also, each message sent by p_i is signed by the signature module just before leaving the process. So, if the sender identity contained in the message is corrupted, this will be discovered by the receiver signature module.

Failure detection module of the class $\diamond\mathcal{S}(bz)$ This module manages the set $suspected_i$. It is devoted to the detection of quiet processes [9]. It can be implemented by a set of time-outs. Upon the receipt of a message sent by process p_k, the signature module of p_i resets the local timer associated with p_k and, if $p_k \in suspected_i$, removes p_k from that set. Then, the message is passed to p_i's arbitrary behavior detection module. When the timer associated with p_k expires, p_k is appended to $suspected_i$. It is important to note that, due to asynchrony, the implementation of the Eventual Weak Accuracy property of $\diamond\mathcal{S}(bz)$ can at best be approximate.

Arbitrary behavior detection module This module receives messages from the failure detection module $\diamond\mathcal{S}(bz)$ and checks if they are properly formed and follow the program specification of the sender. In the affirmative, it passes the message to p_i's certification module. This module maintains a set ($byzantine_i$) of processes it detected to experience at least one byzantine behavior such as duplication, corruption or misevaluation. We say "process p_i declares p_j to be byzantine" at some time, if, at that time, $p_j \in byzantine_i$. Note that differently from the failure detector module, the arbitrary behavior detection module is reliable (i.e., if $p_j \in byzantine_i$, then p_j has experienced an incorrect behavior detected by the arbitrary behavior detection module of p_i). Finally, as for the set $suspected_i$, p_i's consensus module can only read $byzantine_i$. Let us note that $byzantine_i$ cannot be corrupted by process p_i's consensus module, but, this does not prevent p_i's consensus module to misevaluate an expression involving $byzantine_i$ (e.g., $p_j \in byzantine_i$ is evaluated to false by p_i even though it was actually true).

Section 5 is devoted to the implementation of the arbitrary behavior detection modules.

Certification Module This module is responsible, upon the receipt of a message from the arbitrary behavior detection module, for updating the corresponding certificate local variable. It is also in charge to append properly formed certificates to the messages that are sent by p_i.

4 The Vector Consensus Protocol

This section provides a protocol implementing the consensus module of each process. It uses Hurfin-Raynal's protocol (Section 2.4) as a skeleton.

4.1 Local Variables

Each local variable is a way that a byzantine process can use to attack correct processes by corrupting its value. Hence, local variables should be used very rarely and their values should be carefully certified.

- $nb\_current_i$ (resp. nb_next_i) can be replaced by using the cardinality of the certificate $current\_cert_i$ (resp. $next\_cert_i$) which contains properly formed CURRENT (resp. NEXT) votes received in the current round.
- $state_i$ can assume three values (q_0, q_1, q_2). Each state can be identified (when necessary) by using certificates in the following way:
  - $state_i = q_0$: no CURRENT vote has been received by p_i and p_i has not sent a NEXT vote. i.e., $(|current\_cert_i| = 0) \wedge \langle \text{NEXT}(p_i, r_i), cert\rangle_i \notin next\_cert_i$.
  - $state_i = q_1$: a CURRENT vote has been received by p_i and p_i has not sent a NEXT vote: $(|current\_cert_i| \geq 1) \wedge \langle \text{NEXT}(p_i, r_i), cert\rangle_i \notin next\_cert_i$.
  - $state_i = q_2$: p_i has sent a NEXT vote: $\langle \text{NEXT}(p_i, r_i), cert\rangle_i \in next\_cert_i$.
- rec_from_i can be replaced by the variable REC_FROM_i using certificates in the following way:
  $REC\_FROM_i \equiv \{p_\ell | \langle \text{NEXT}(p_\ell, r_\ell), cert\rangle_\ell \in next\_cert_i \vee$
  $\langle \text{CURRENT}(p_\ell, est\_vect_\ell, r_\ell), cert\rangle_\ell \in current\_cert_i\}$

The predicate *change_mind*
For the sake of brevity, let *change_mind* denote the following predicate:

$$(|current\_cert_i| \geq 1) \wedge \langle \text{NEXT}(p_i, r_i), cert\rangle_i \notin next\_cert_i \wedge |REC\_FROM_i| \geq (n - F))$$

This predicate corresponds to the predicate $(state_i = q_1) \wedge (\mid rec\_from_i \mid > n/2)$ used in the line 17 of the protocol shown in Figure 1.

Note that the only variables that cannot be replaced by the use of certificates are the round number (r), the coordinator (c, which directly depends on r) and the current estimates (est_vect). Then, their values must be authenticated by certificates as explained below.

4.2 Certificates Attached to Messages

Four types of messages are exchanged, namely, INIT, CURRENT, NEXT and DECIDE. Each time a message m (INIT, CURRENT, NEXT) is received by the certification module, m is appended to a local variable keeping the corresponding certificates (*est_cert*, *current_cert* or *next_cert*, respectively). These statements are depicted inside a box in the protocol described in Figure 3 (see line 8 for INIT, line 18 for CURRENT and line 28 for NEXT).

Each time a process p_i sends a message m, according to the type of m, an appropriate certificate is associated with m by the certification module. This certificate depends on the protocol statement that issues the sending of m. This certificate will be used by the arbitrary behavior detection module of the receiver to check the proper sending of m.

Certifying initial values for Vector Consensus For each process p_i, the first problem lies in obtaining a vector of proposed values (certified vector) that verifies the Vector Validity property (Section 3). This certified vector will then be used as the value proposed by p_i to the consensus protocol.

This certification procedure is similar to the one proposed by Doudou and Schiper in [4]. It is described in the protocol of Figure 3 (line 4 to line 9). Each process initially broadcasts its value v_i and then waits for $(n - F)$ values from other processes. Each time p_i receives a value, the message is added to est_cert_i.

Exiting from the initial while loop (lines 6-9) we say that "*est_cert_i is well-formed with respect to a value est_vect_i*" if the following conditions are satisfied:

- $|est\_cert_i| = (n-F)$ (otherwise process p_k has either omitted to execute the receipt of line 7 or misevaluated the condition of line 6), and
- The value est_vect_i is correct with respect to the $(n - F)$ INIT messages contained in est_cert_i. Otherwise process p_i either has omitted to execute the update of est_vect_i (line 8), or has corrupted its value (intuitively, this means that those $(n - F)$ messages "witness" that est_vect_i is a correct value, because $n - F \geq n - C$).

Certifying estimate values The initial value of est_vect_i (obtained when exiting from lines 4-9) is certified by est_cert_i, as explained above. This variable can then take successive values: it can be updated at most once per round (line 18) due to the delivery of the first CURRENT message received during this round (line 15). When this occurs, the certificate est_cert_i is also updated. Since this message is properly formed, its certificate $cert_k$ contains a correct certificate est_cert_k (*i.e.*, a certificate well-formed with respect to the value est_vect_k contained in the CURRENT message). During a round, est_cert_i is said to be "*well-formed with respect to est_vect_i*" if the value est_vect_i is the value included in the $(n - F)$ messages contained in est_cert_k. Otherwise, it means that process p_i has either omitted to execute the update of est_vect_i (line 18), or corrupted its value.

From round $r-1$ to round r A process progresses from round $r-1$ to round r when the predicate of line 14 is false. When a new round r starts, we say that $next\_cert_i$ *is well-formed with respect to* $r - 1$ if the following two conditions are satisfied:

- $|next\_cert_i| = (n - F)$ and $r > 1$ (otherwise process p_i has misevaluated the condition of line 14).
- The value $r - 1$ is consistent with respect to the information in the $(n - F)$ NEXT messages contained in est_cert_i (i.e., all messages refer to round $r - 1$. Otherwise process p_k has corrupted the value of r at line 11).
- If $r = 1$, then $next\_cert_i = \emptyset$ (Otherwise either p_i has corrupted r or c (line 11) or has misevaluated the condition of line 12).

4.3 Protocol Description

The text of the protocol is presented in Figure 3. In the following, a *vote* means a message CURRENT or NEXT, and a *valid vote* means a properly signed and formed vote.

- At the beginning of a round r, the current coordinator p_c proposes its estimate est_vect_c to become the decision value by broadcasting a CURRENT vote carrying this value (line 12). This vote is certified by $est\_cert_c \cup next\_cert_c$. est_cert_c is used to certify the value proposed by the coordinator est_vect_c. i.e.,

est_cert_c must be well formed wrt est_vect_c. $next\_cert_c$ is used to certify the value of the current round r (i.e., $next\_cert_c$ must be well formed wrt $r-1$).

- When p_i receives the first CURRENT valid vote while in state q_0 (line 17). It relays a CURRENT vote (line 19) by using the valid vote CURRENT just received as a certificate. This certificate contains $est\_cert_c \cup next\_cert_c$ used to certify r and est_vect_c (as above).

- If, while it is in the initial state q_0, p_i suspects the current coordinator, it broadcasts a NEXT vote (line 25) and moves to q_2. This vote is certified by $est\_cert_i \cup current\_cert_i \cup next\_cert_i$. Those certificates ($current\_cert_i$ and $next\_cert_i$) will be used by the arbitrary behavior detection module of the receiver to decide whether p_i has misevaluated or not the sending condition (at line 24). Moreover, as NEXT votes can also be sent at lines 29 and 31, est_cert_i is used to allow the receiver to determine the condition that has triggered the NEXT vote it receives.

- When the predicate at line 29 becomes true, in order to avoid a deadlock, process p_i broadcasts a NEXT vote to favor the transition to the next round (line 30). This vote is certified by $current\_cert_i \cup next\_cert_i$. $current\_cert_i$ and $next\_cert_i$ are used to certify the non-misevaluation of the predicate $change\_mind$.

- When a process progresses from round r to round $r+1$ it issues a NEXT vote if it did not do it in the **while** loop. These NEXT votes are used to prevent other processes from remaining blocked in round r (line 32). This vote is certified by $next\_cert_i$ which will allow a receiver to check the correct evaluation of the condition at line 14 by verifying if $next\_cert_i$ is well formed wrt r.

4.4 Taking a Decision

A process decides a value est_vect_c at round r either when it has received $(n-F)$ valid CURRENT votes (lines 21-22) or when it receives a properly signed and formed DECIDE message from another process (lines 2-3). In the first case the process authenticates its decision by using a *well-formed* $current\_cert_i$ as a certificate (line 22), in the second case the message (with the same certificate) is relayed to the other processes (line 3).

We say that $current\_cert_i$ is well formed wrt r and est_vect if the following conditions are satisfied:

- $|current\_cert_i| = (n-F)$ (otherwise process p_i misevaluated the condition of line 21).

- the certificate of each message in $current\_cert_i$ contains a est_cert_c well formed with respect to est_vect.

- the certificate of each message in $current\_cert_i$ contains a $next\_cert$ well formed with respect to r.

4.5 Summary

Thanks to the verifications made by the Signature verification module and the arbitrary behavior detection module, we have, for every correct process:

- Every accepted INIT(p_k, v_k) message has a correct field p_k (signature module).

- Every accepted CURRENT(p_k, r, est_vect_k) message has correct fields p_k (Signature verification module), r (certificate $next\_cert_k$) and est_vect_k (certificate est_cert_k contained in $current\_cert_k$).

- Every accepted NEXT(p_k, r) message has correct fields p_k (signature module) and r (certificate $next\_cert_k$).

- Every accepted DECIDE(p_k, est_vect_k) message has correct fields p_k (Signature verification module), and est_vect_k (certificates contained in $current\_cert_k$).

5 The Arbitrary Behavior Detection Module

The arbitrary behavior detection module of process p_i is composed of a set of finite state automata, one for each process. The module associated with p_i is depicted in Figure 4. This module detects if a given process p_k has suffered a byzantine fault. This detection is carried out in two steps. (i) It is first checked if a certificate of a message is well formed with respect to the value carried by this message. It is then (ii) checked if the type of the message follows the program specification (i.e., if it has been sent at the right time). If one of these steps

```
function consensus(v_i)

(1) [next_cert_i ← ∅] ; [est_cert_i ← ∅] ; r_i ← 0;
    cobegin
(2) || upon receipt of a properly signed and formed ⟨ DECIDE(p_k, est_vect_k), cert_k⟩_k
(3)     send ⟨DECIDE(p_i, est_vect_k), cert_k⟩_i to Π;  return(est_vect_k)

(4) || for k ← 1 to n do est_vect_i[k] ←  null endfor;
(5)     send ⟨INIT(p_i, v_i), ∅⟩_i to Π;
(6)     while  |est_cert_i| ≠ (n − F) do
(7)         wait receipt of ⟨INIT(p_k, v_k), cert_k⟩_k;
(8)         [est_cert_i ← est_cert_i ∪ ⟨INIT(p_k, v_k), cert_k⟩_k] ; est_vect_i[k] ← v_k
(9)     end while

(10)    loop % on a sequence of asynchronous rounds %
(11)        c ← (r_i mod n) + 1; r_i ← r_i + 1;
(12)        if  (i = c) then send ⟨CURRENT(p_i, r_i, est_vect_i), est_cert_i ∪ next_cert_i⟩_i to Π endif; % q_0 → q_1 for i = c %
(13)        [next_cert_i ← ∅] ; [current_cert_i ← ∅] ;
(14)        while (|next_cert_i| ≤ (n − F)) do
(15)            upon receipt of a properly signed and formed ⟨CURRENT(p_k, r_i, est_vect_k), cert_k⟩_k
(16)                [current_cert_i ← current_cert_i ∪ ⟨CURRENT(p_k, r_i, est_vect_k), cert_k⟩_k] ;
(17)                if (|current_cert_i| = 1) then [est_cert_i ← cert_k] ; est_vect_i ← est_vect_k endif;
(18)                if (|current_cert_i| = 1) ∧(⟨NEXT(p_i, r_i), cert_i⟩_i ∉ next_cert_i)  ∧  (i ≠ c) % q_0 → q_1 for i ≠ c %
(19)                                    then send ⟨CURRENT(p_i, r_i, est_vect_i), current_cert_i⟩_i to Π endif;
(20)                if (|current_cert_i| = (n − F)) then
(21)                        send ⟨DECIDE(p_i, est_vect_i), est_cert_i⟩_i to Π;  return(est_vect_i) endif

(22)            upon (p_c ∈ (suspected_i ∨ byzantine_i))
(23)                if  ((|current_cert_i| = 0) ∧ ⟨NEXT(p_i, r_i), cert_i⟩_i ∉ next_cert_i)
(24)                    then  send ⟨NEXT(p_i, r_i), current_cert_i ∪ next_cert_i ∪ est_cert_i⟩_i to Π % q_0 → q_2 %
(25)                endif

(26)            upon receipt of a properly signed and formed ⟨NEXT(p_k, r_i), cert_k⟩_k
(27)                [next_cert_i ← next_cert_i ∪ ⟨NEXT(p_k, r_i), cert_k⟩_k]

(28)            upon ( change_mind ) % q_1 → q2 %
(29)                send ⟨NEXT(p_i, r_i), current_cert_i ∪ next_cert_i⟩_i to Π
(30)        endwhile

(31)        if (⟨NEXT(p_i, r_i), cert_i⟩_i ∉ next_cert_i) then send ⟨NEXT(p_i, r_i), next_cert_i⟩_i to Π  endif % q_0/q_1 → q_2 %
(32)    endloop
    coend
```

Figure 3: The Consensus Module of Process p_i

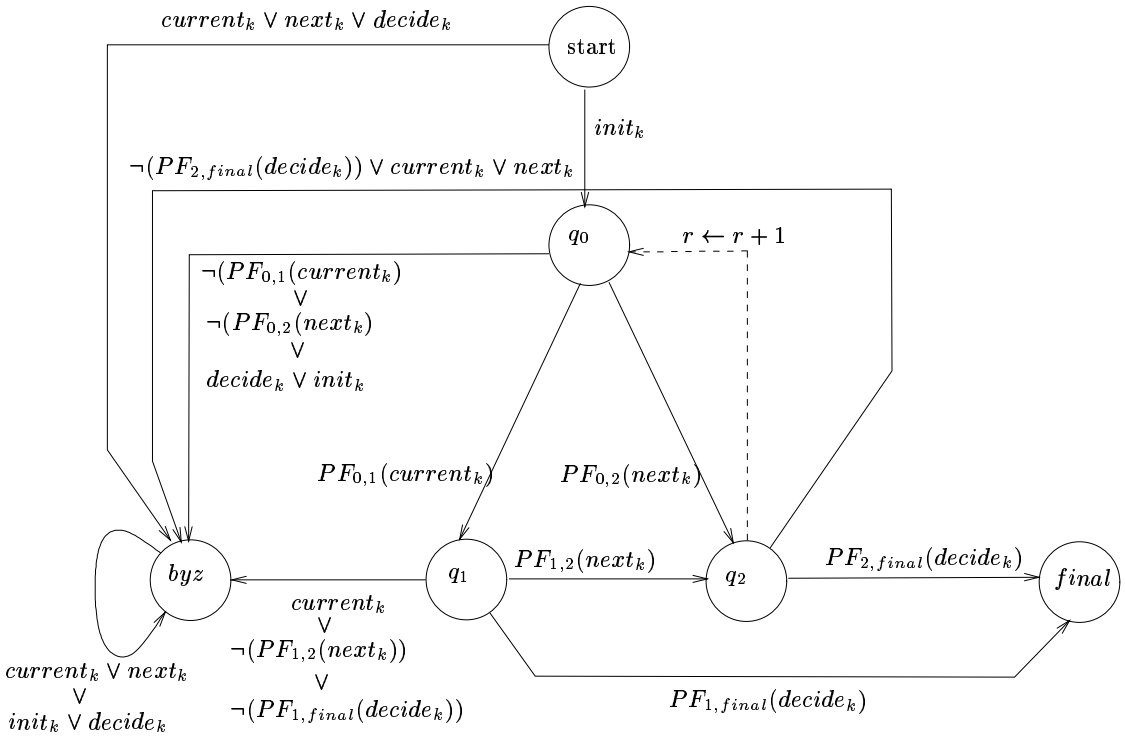

Figure 4: Finite state automaton of the p_i's arbitrary behavior detection module to monitor process p_j.

detects something wrong, the automaton falls in the state *byz* which means p_k is byzantine, furthermore, the message is discarded. Otherwise, the message is passed to the consensus module of process p_i.

The automaton of p_i related to p_k represents the view p_i has, during the current round r_i, on the behavior of p_k with respect to its automaton (made of states q_0, q_1 and q_2, see Sections 2.4 and 4) during the same round. It evaluates a condition each time a message has arrived from p_k. According to the result of this evaluation, it moves from its current state a to another state b. The condition is composed of the following predicates. $PF_{a,b}(type\_of\_message_k)$ returns true if the message is properly formed. The predicate $type\_of\_message_k$ is true if a message of that type has been received from p_k.

Automaton States The automaton of process p_i related to a process p_k is composed of 6 states. Three states are related to a single round (as the ones described in the previous section: q_0, q_1, q_2), one is the initial state (*start*), one is the final state ($final$) and one is the state declaring p_k is byzantine (*byz*). Note that the predicate $p_k \in byzantine_i$ is true if the automata related to p_k of process p_i is in state *byz*. It is false otherwise.

Automaton Transitions

- *Transition start* $\rightarrow q_0$. It occurs as soon as an INIT message is received from p_k i.e., $init_k$ becomes true.
- *Transition start* $\rightarrow byz$. It occurs when p_i receives any other message from p_k (e.g., CURRENT, NEXT or DECIDE) before INIT. Due to the FIFO property of channels, p_i concludes that p_k is byzantine as it omitted to execute line 5 and moves to state *byz*.
- *Transition* $q_0 \rightarrow q_1$. The arbitrary behavior detection module of process p_i receives $\langle \text{CURRENT}(p_k, r_i, est\_vect_k), cert_k \rangle_k$ from p_k. We have two cases:

  Case 1. p_k is the coordinator (the message was sent by p_k from line 12). i.e., $cert_k = est\_cert_k \cup next\_cert_k$. The predicate $PF_{0,1}(current_k)$ returns true if:

  1. est_cert_k is well-formed with respect to a value est_vect_k, and
  2. $next\_cert_k$ is well-formed with respect to the round number $r_i - 1$.

Case 2. p_k is not the coordinator (the message was sent by p_k from line 19). i.e., $cert_k = current\_cert_k$. According to the protocol, the certificate must include the message $\langle \text{CURRENT}(p_c, est\_vect_c, r_i), cert\rangle_c$ in order to be properly formed (line 16), then, once extracted this message from the certificate, all tests of case 1 can be executed on $est\_cert_c \cup next\_cert_c$.

- *Transition* $q_0 \rightarrow q_2$. Upon the arrival of a $\langle \text{NEXT}(p_k, r_i), cert_k\rangle_k$ at the arbitrary behavior detection module, we have two cases:

  - If $cert_k$ does not contain any INIT message, it was sent by p_k from line 32 where $cert_k = next\_cert_k$. In such a case $PF_{0,2}(next)$ is true if $cert_k$ is well formed wrt $r_i - 1$. Otherwise there was a misevaluation of the condition at line 14.
  - If $cert_k$ contains at least one INIT message, the NEXT message was sent by p_k from line 25 with $cert_k = current\_cert_k \cup next\_cert_k \cup est\_cert_k$. This sending is guarded by the conditions at line 24. $PF_{0,2}(next)$ is true, if by using the information contained in $cert_k$, est_cert_k is well-formed with respect to a value est_vect_k and the conditions at line 24 (i.e., $((|current\_cert_k| = 0) \wedge \langle \text{NEXT}(p_k, r_i), cert\rangle_k \notin next\_cert_k))$ can be evaluated to true.

  Remark Note that even though $PF_{0,2}(next)$ is true, the process p_k could be byzantine as it could misevaluate the predicate $p_c \in (suspected_i \vee byzantine_i)$ used at line 23 and no information about that fault can be found in the certificate. One solution to this problem is to assume $p_c \in (suspected_i \vee byzantine_i)$ *cannot be misevaluated* by a process. The other is to say that, as at most F processes are byzantine, then even though all of them misevaluate that predicate and generate false NEXT messages, this does not prevent other processes to get a decision due to conditions of line 14. *End of remark.*

- *Transition* $q_0 \rightarrow byz$. Upon the arrival of a message, p_i declares p_k byzantine, if one of the following four conditions is true: (i) $PF_{0,1}(current)$ is false (in the case of the receipt of a CURRENT message); or (ii) $PF_{0,2}(next)$ is false (in the case of the receipt of a NEXT message); or (iii) the message is of type DECIDE; or (iv) the message is of type INIT. In the case (iii), as channels are FIFO, a NEXT or DECIDE message has been omitted by p_k. In the case (iv) process p_k executed twice the statement at line (5).

- *Transition* $q_1 \rightarrow q_2$. Upon the arrival of a $\langle \text{NEXT}(p_k, r_i), cert\rangle_k$ from p_k at the arbitrary behavior detection module, it executes the following tests on the certificate $cert_k$. The NEXT message was sent by p_k from line 30 with $cert_k = current\_cert_k \cup next\_cert_k \cup est\_cert_k$. That sending is guarded by the condition at line 29.

  $PF_{1,2}(next)$ is true if, using the information contained in $cert_k$, the predicate *change_mind* can be evaluated to true. As above remark the fact that $PF_{1,2}(next)$ is true does not imply that the sender process is correct, as the condition at line 29 includes the predicate $p_c \in (suspected_i \vee byzantine_i)$ that can be misevaluated by p_k without being detected.

- *Transition* $q_1 \rightarrow final$ *and transition* $q_2 \rightarrow final$. These transitions occur upon the arrival of a message $\langle \text{DECIDE}(p_k, est\_vect_k), cert\rangle_k$ at the arbitrary behavior detection module. $PF_{1,2}(decide)$ (resp. $PF_{1,final}(decide)$) is true if $cert_k$ is well formed with respect to est_vect_k and r_i.

- *Transition* $q_1 \rightarrow byz$. Upon the receipt of a message, p_i declares p_k byzantine, if one of the following three conditions is true: (i) $PF_{1,2}(next_k)$ is false (in the case of the receipt of a NEXT message); or (ii) $PF_{1,final}(decide_k)$ is false (in the case of the receipt of a DECIDE message); or (iii) the message is of type CURRENT. In the latter case, as channels are FIFO, a NEXT or DECIDE message should be received by p_i from p_k, then p_k had a byzantine failure.

- *Transition* $q_2 \rightarrow byz$. Upon the receipt of a message, p_i declares p_k byzantine, if one of the following three conditions is true: (i) $PF_{2,final}(decide_k)$ is false (in the case of the receipt of a DECIDE message); or (ii) the message is of type CURRENT; or (iv) the message is of type NEXT. In the latter two cases, as channels are FIFO only a DECIDE message should be received by p_i from p_k, if a then message of type CURRENT orNEXT is received from p_k, it had a byzantine failure.

- *Transition* $byz \rightarrow byz$. Once p_k is declared *byzantine* it remains in the state *byz* whatever type of message is received.

- *Transition $q_2 \rightarrow q_0$ (denoted by a dotted line).* A process p_k in p_2 either decides a value, and then moves to *final* or moves to q_0 (and then to the successive round) in a finite time. The fact that p_k moved to q_0 can be detected if the successive message received by the process from p_k is either a NEXT or a CURRENT message related to round $r+1$.

6 Correctness Proof

This section proves that the previous protocol satisfies the Termination, Vector Validity and Agreement properties. This proof assumes that:

- **H1**: There are at most F processes that do not behave correctly, with $F \leq \min(\lfloor \frac{n-1}{2} \rfloor, C)$.
- **H2**: The underlying failure detection module belongs to the class $\diamond\mathcal{S}(bz)$, *i.e.*, it satisfies:

  H2.1: Strong Completeness (Eventually, every quiet process is permanently suspected by every correct process). And,

  H2.2: Eventual Weak Accuracy (There is a time after which some correct process is never suspected by any correct process).
- **H3**: The underlying arbitrary behavior detection module is reliable (if $p_j \in byzantine_i$ then p_j has experienced at least one byzantine behavior).
- **H4**: Communication channels are FIFO.

6.1 Vector Validity

Lemma 1 *Eventually, every correct process builds a vector est_vect_i such that $est\_vect_i[i] = v_i$ and $\forall k \neq i$: $est\_vect_i[k] = v_k$ or $est\_vect_i[k] = null$.*

Proof From assumptions **H1** and **H4**, the while loop eventually terminates (line 9) and the values contained in est_vect_i are either *null* line 4) or set at line 8. □

Lemma 2 *During any round r, for any process p_i, we have: $|current\_cert_i| > 0 \Rightarrow est\_vect_i = est\_vect_c$, where p_c is the current coordinator. In particular, $est\_vect_i[i] = v_i$ or $est\_vect_i[i] = null$.*

Proof **case** $i = c$. est_vect_c is updated with the value est_vect_c at most once in the current round, at line 18. **case** $i \neq c$. est_vect_i is updated at most once per round, at line (18), when p_i receives for the first time a valid vote CURRENT. Let p_k be the sender of this vote.

- If $k = c$ then $est\_vect_i \leftarrow est\_vect_c = est\_vect_c$
- If $k \neq c$ let us examine the sequence of votes CURRENT leading to the update of line 18:
  - Every process p_j ($j \neq c$) sending a vote CURRENT (at line 19) has received a valid vote $\langle \text{CURRENT}(-, vect, -), - \rangle$ and has performed line 18: $est\_vect_j \leftarrow vect$
  - Only p_c can initiate such a sequence of CURRENT messages.

  From this follows that $est\_vect_i = est\_vect_k = \ldots = est\_vect_j = \ldots = est\_vect_c$.

Thus, in particular, $est\_vect_i[i] = est\_vect_c[i]$ and this value is either v_i or *null*, from Lemma 1. □

Lemma 3 *No process can build two different initial certified vectors.*

Proof During the initial phase (lines 6-9) a process p_i can receive at most one INIT message from each process p_k. In fact, INIT messages are signed and the arbitrary behavior detection module filters out duplicate messages. Suppose that p_i builds two different initial certified vectors est_vect_i and $est\_vect'_i$ (with $est\_vect_i \neq est\_vect'_i$). In that case, p_i has two certificates est_cert_i and $est\_cert'_i$ well-formed respectively wrt est_vect_i and $est\_vect'_i$. In particular, $|est\_cert_i| = |est\_cert'_i| = (n - F)$. Each of these certificates contains $(n - F)$ identities of processes, namely the senders of the $(n - F)$ signed INIT messages forming these certificates. Let X (resp. X')

denote the set of process identities belonging to $cert_i$ (resp. $cert'_i$). By assumption, $| X |=| X' |= (n-F)$. On the other hand, $| X \cap X' |=| X | + | X' | - | X \cup X' |$, and thus $| X \cap X' | \geq 2(n-F) - n = n - 2F$. From assumption **(H1)**, we get $n - 2F \geq 1$, which means $X \cap X' \neq \emptyset$. Thus, p_i has received *two* INIT messages from a same sender. A contradiction. □

Lemma 4 *A correct process sending a message DECIDE (whose initial sending has been initiated at round r, line 21) decides est_vect_c (where p_c is the coordinator of round r).*

Proof A process decides at line 3 or at line 22.

- If p_i decides at line 22, then $|current\_cert_i| > 0$ and thus, from Lemma 2, p_i decides $est\_vect_i = est\_vect_c$. Before deciding, it sends a message DECIDE with the value est_vect_c to all other processes.
- If p_i decides at line 3, it decides the value contained in the properly signed and formed message DECIDE. This message contains est_vect_c.

□

Theorem 1 *If a correct process decides, it is on a vector v of size n, satisfying Vector Validity with at least $\alpha = n - 2F \geq 1$ entries from correct processes.*

Proof Every process builds a vector (Lemma 1), containing $(n-F)$ values from different processes (lines 6-9) and F null values (line 4). If a process decides at round r, then it decides est_vect_c (Lemma 4). If p_c has sent a CURRENT vote, est_cert_c is well-formed wrt est_vect_c. In particular, est_vect_c is correct wrt the $(n-F)$ INIT messages contained in $cert\_init_c$. So, est_vect_c contains $(n-F)$ initial values of different processes. Moreover, from Lemma 1, $est\_vect_c[i] = v_i$ or $est\_vect_c[i] = null$.

Let $nb\_correct_c$ (resp. nb_faulty_c) denote the number of non-null entries of est_vect_c that are initial values from correct (resp. non correct) processes. We have $nb\_correct_c + nb\_faulty_c = (n-F)$. But $nb\_faulty_c \leq F$ and thus $nb\_correct_c \geq n - 2F$. From assumption **(H1)**, we have $n - 2F \geq n - 2\lfloor \frac{n-1}{2} \rfloor \geq 1$.

□

6.2 Termination

Lemma 5 *If a correct process decides, then eventually every correct process will decide.*

Proof Let p_i deciding at line 3 or at line 22. In both cases, p_i sends a properly signed and formed message DECIDE to every other process. From assumption **H4**, every correct process that did not yet decide eventually receive this message (line 2) and decides (line 3). □

Lemma 6 *If no process decides during any round $r' \leq r$, then all correct processes start round $r+1$.*

Proof The proof is by contradiction. Suppose no process has decided in any round $r' \leq r$, where r is the smallest round number in which some correct process blocks forever in the **while** loop (lines 14-31). Let us note that no correct process has received a DECIDE message (otherwise, it would execute lines 2-3 and decide). In the following proof, we use the following notations, recalling the "hidden" states q_0, q_1, q_2 of the processes (cf. page 10):

state $q_0 \equiv (|current\_cert_i| = 0) \wedge \langle \text{NEXT}(p_i, r_i), cert \rangle_i \notin next\_cert_i$,
state $q_1 \equiv (|current\_cert_i| \geq 1) \wedge \langle \text{NEXT}(p_i, r_i), cert \rangle_i \notin next\_cert_i$,
state $q_2 \equiv \langle \text{NEXT}(p_i, r_i), cert \rangle_i \in next\_cert_i$.

1. Note first that any process starting round r is in state q_0 before entering the while loop (lines 11 and 13). Firstly, it is shown that a correct process p_i can not remain in state q_0 during round r. This follows from:

   i) Either p_i suspects p_c and moves to q_2 (lines 23-26). This is due either to the underlying failure detector (possibly mistaken), or because p_c is detected as Byzantine by the underlying arbitrary behavior detection module.

ii) Or p_i never suspects p_c. Due to assumption H2.1 (Strong Completeness), this means that p_c is not quiet for p_i: eventually it broadcasts a CURRENT vote. Due to assumption **H4**, p_i receives at least one CURRENT valid vote (from p_c or from another process) and moves to q_1 (line 16).

2. A correct process can not remain in state q_1.
By contradiction: suppose that a correct process p_i remains in state q_1. From the previous point, any correct process either sends a CURRENT vote and moves to q_1 (Case ii), or sends a NEXT vote and moves to q_2 (Case i). It follows from assumption **H1** that any correct process will receive at least $(n - F)$ valid votes and thus eventually $|REC\_FROM_i| \geq (n - F)$. Thus p_i eventually satisfies *change_mind*. Hence, the condition stated at line 29 eventually becomes true: p_i issues a NEXT vote and moves to q_2.

3. It follows from the two previous points that all correct processes move to q_2 and send a NEXT vote. Consequently, any correct process p_i receives at least $(n - F)$ NEXT valid votes, and thus proceeds to round $r + 1$. A contradiction.

□

Lemma 7 *During a round r, if no process suspects the coordinator (or, equivalently, no process moves from state q_0 directly to state q_2) within the* **while** *loop, then no process sends a NEXT vote (or equivalently, no process moves to q_2).*

Proof From the lemma assumption, no process executes lines 23-26 (going to state q_2). Suppose that, during round r, a process sends a NEXT vote. So, there exists a process that has sent a NEXT vote before receiving a NEXT valid vote. Such a process p_i executes line 30 or line 32.

- p_i executes line 30.
This is possible only because *change_mind* has become true (line 29), *i.e.*, p_i has received a (CURRENT or NEXT) valid vote from at least $(n - F)$ processes. If one of these votes is a NEXT, this contradicts the fact that p_i has sent a NEXT vote before receiving a NEXT valid vote. So, all these valid votes are CURRENT. Thus, p_i has decided at line 22 when it received the last valid vote making true the condition $|current\_cert_i| = (n - F)$. Consequently, as p_i has executed a **return** statement it will never execute line 30. Contradiction.

- p_i executes line 32.
To send a NEXT vote at line 32, p_i has terminated its **while** loop. So, it has $|next\_cert_i| > F$. It follows that p_i has received NEXT valid votes before executing line 32. Contradiction.

□

Theorem 2 *Every correct process eventually decides some value.*

Proof Let us consider the two following cases.

1. A correct process p_j decides (at line 22). From Lemma 5, every correct process eventually decides.

2. No process decides. We will show there is a contradiction. There is a time t after which (due to assumption **H2.2**, Eventual Weak Accuracy) there is a correct process that is no longer suspected (let p_j be this process). Let r be the first round that occurs after t and which is coordinated by p_j (due to Lemma 6, such a round does exist since no process decides).
   - The coordinator p_j sends a CURRENT vote to all processes.
   - As, by assumption, the current coordinator p_j is not suspected, no process p_i executes lines 23-26. Consequently, no process moves directly from q_0 to q_2 within the **while** loop.
   - From Lemma 7, we conclude that no process sends a NEXT vote, *i.e.*, no process moves to q_2.
   - A process entering the loop while (line 14) cannot exit this loop, because $|next\_cert_i|$ remains equal to 0.
   - From assumptions **H1** and **H4**, each correct process p_i will receive at least $(n - F)$ CURRENT valid votes. As no process sends a NEXT vote, it follows that any correct process p_i will necessarily decide at line 22. This contradicts the assumption that no process decides.

□

6.3 Agreement

Lemma 8 *If a process decides v and broadcasts a properly signed and formed DECIDE message at line 21 of round r, then all correct processes p_j that start round $r+1$ do so with $est\_vect_j = v$.*

Proof

1. As p_i broadcasts a message DECIDE labeled with the round number r, some process p_k (possibly p_i) has processed the line 21 during round r and, consequently, $|current\_cert_k| = (n-F)$ **(R1)**. Moreover, from Lemma 4, the decision value v is equal to est_vect_c.

2. Let us consider the three following sets of processes (related to round r):
   – X_1^r = { processes that moved from q_0 to q_1 and did not move to q_2 }
   – X_2^r = { processes that moved from q_0 directly to q_2 }
   – X_3^r = { processes that moved from q_0 to q_1 and then to q_2 }
   Note that these sets are disjoint. They include processes that have possibly behaved incorrectly after moving from q_0 to another state. Moreover, we have $|X_1^r| + |X_2^r| + |X_3^r| \leq n$ **(R2)**.

3. Let $sent\_to\_pk$ be the number of processes that sent a CURRENT vote to p_k. As the number of CURRENT votes sent to a p_k is greater than or equal to the number of CURRENT valid votes received by p_k, we have $sent\_to\_pk \geq |current\_cert_k|$ **(R3)**.

4. All processes belonging to X_3^r have sent a CURRENT vote to p_k at line 19 (when they moved from q_0 to q_1). Moreover, all processes belonging to X_1^r and which executed all of line 19 have sent a CURRENT vote to p_k. Some processes belonging to X_1^r and which have partially executed line 19 have also sent a CURRENT vote to p_k. Finally, processes in X_2^r have not sent a CURRENT vote. It follows that $sent\_to\_pk \leq |X_1^r| + |X_3^r|$ **(R4)**.

5. From **(R1)** and **(R3)**, we conclude $sent\_to\_pk \geq (n-F)$ **(R5)**.

6. From **(R5)** and **(R4)**, we conclude $|X_1^r| + |X_3^r| \geq (n-F)$ **(R6)**.

7. From **(R6)** and **(R2)**, we conclude $|X_2^r| \leq F$ **(R7)**.

The proof is now by contradiction. Suppose that p_i decides (and consequently **(R7)** holds), and that there is a process p_j which enters round $r+1$ with $est\_vect_j \neq v$ (*i.e.*, $est\_vect_j \neq est\_vect_c$). Let us consider the value $|current\_cert_j|$ just before p_j leaves round r. There are two cases.

1. $|current\_cert_j| > 0$. From Lemma 2, we have $est\_vect_j = est\_vect_c$, a contradiction.

2. $|current\_cert_j| = 0$. In this case, since, according to the lemma assumption, p_j proceeds to the next round, we have $|next\_cert_j| > F$ **(R8)**, at the end of round r. Combining **(R7)** and **(R8)**, we get $|next\_cert_j| > |X_2^r|$ **(R9)**.

   Since NEXT votes are only sent by processes belonging to $X_2^r \bigcup X_3^r$, and as (by definition) $X_2^r \bigcap X_3^r = \emptyset$, from **(R9)** we conclude that p_j received at least one NEXT valid vote from a process $p_\ell \in X_3^r$. As p_ℓ belongs to X_3^r:

   - $\ell \neq c$ and p_ℓ first passed through state q_1. So, it processed lines 17-19, from which we conclude $|current\_cert_\ell| > 0$ and, from Lemma 2, $est\_vect_\ell = est\_vect_c$. In particular, p_ℓ has sent a CURRENT vote to p_j.
   - p_ℓ then moved from q_1 to q_2. So, it necessarily sent the NEXT vote (received by p_j) at line 30.

   From the assumption **H4** (channels are FIFO), p_j has received from p_ℓ the $\langle \text{CURRENT}(p_\ell, r_\ell, est\_vect_\ell), current\_cert_\ell \rangle_\ell$ valid vote first (line 15), then the $\langle \text{NEXT}(p_\ell, r_\ell), current\_cert_\ell \cup next\_cert_\ell \rangle_\ell$ (line 27). Upon the receipt of the CURRENT valid vote, the condition of line 17 holds and p_j has updated est_vect_j to $est\_vect\_\ell = est\_vect\_c$. Upon the receipt of the NEXT valid vote, no update of est_vect_j occurred. So, just before p_j leaves round r, $est\_vect_j = est\_vect_c$. A contradiction.

□

Theorem 3 *No two correct processes decide different values.*

Proof Let p_i and p_j two correct processes that decide. There are two cases:

1. Both p_i and p_j decide at the same round r (both execute the line 21 during the round r). Let est_vect_c denote the certified initial vector built by the coordinator p_c of round r (Lemma 1). From Lemma 4, a process that decides at round r decides the value est_vect_c. From Lemma 3, the coordinator cannot build two different certified initial vectors. So, p_i and p_j decide the same value est_vect_c.

2. p_i decides v at round r and p_j decides at round $r' > r$. From Lemma 8, every correct process p_k that starts round $r+1$ does so with $est\_vect_k = v$. In other words, the only possible estimate value for a correct process participating in any round $r' \geq r+1$, is now $v = est\_vect_c$. So, whatever the coordinator of round r', due to Lemma 4, the value decided by p_j will be $v = est\_vect_c$

□

7 Conclusion

This paper has investigated a new approach to solve the consensus problem in asynchronous distributed systems where processes are prone to byzantine failures. This approach is based on the use of several underlying modules. One is the unreliable failure detector module of the class $\diamond\mathcal{S}(bz)$ [9]. This module (whose implementation can only approximate the specification) copes with quiet processes. Another module (arbitrary behavior detection module) is new. It copes with arbitrary byzantine process behavior.

The paper has presented a protocol that solves the (vector) consensus problem in a byzantine system. The proposed protocol is resilient to F faulty processes, $F \leq \min(\lfloor \frac{n-1}{2} \rfloor, C)$ (where C is the maximum number of faulty processes the underlying certification mechanism can tolerate). As it ensures the Vector Validity property, it can be used as a building block to solve other agreement problems, such as Atomic Broadcast [4]. The paper has also presented an implementation of the arbitrary behavior detection module. This implementation is based on a set of finite state automata. An additional interest of the proposed protocol lies in its systematic use of certificates associated with messages. This facilitates the implementation of the arbitrary behavior detection module. Moreover, reducing the number of local variables, also reduces their incorrect management by faulty processes.

The design of protocols able to cope with processes exhibiting a byzantine behavior is a real practical challenge [1]. Malicious attacks and software errors are increasingly common and (unfortunately) go along with the development of applications on top of Internet. The architectural approach and the consensus protocol developed in this paper constitute endeavors to take up with this challenge.

Acknowledgments

The authors are grateful to Michel Hurfin and Achour Mostefaoui with whom they had very interesting discussions on the consensus problem.

References

[1] Castro M. and Liskov B., Practical Byzantine Fault Tolerance. *Proc. 3rd Symposium on Operating Systems Design and Implementation (OSDI'99)*, New Orleans (LO), pp. 173-186, February 1999.

[2] Chandra T. and Toueg S., Unreliable Failure Detectors for Reliable Distributed Systems. *Journal of the ACM*, 34(1):225–267, March 1996.

[3] Chandra T., Hadzilacos V. and Toueg S., The Weakest Failure Detector for Solving Consensus. *Journal of the ACM*, 43(4):685–722, July 1996.

[4] Doudou A. and Schiper A., Muteness Failure Detectors for Consensus with Byzantine Processes, *Brief announcement in Proc. 17th ACM Symposium on Principles of Distributed Computing*, p. 315, 1998. Expanded version in: *Tech. Report 97/30*, EPFL, Lausanne, Switzerland, 1997.

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

[6] Hurfin M. and Raynal M., A Simple and Fast Asynchronous Consensus Protocol Based on a Weak Failure Detector. To appear in *Distributed Computing*, 12(4), 1999. *Tech. Report 1118*, IRISA, France, 1997.

[7] Hurfin M., Raynal M. and Tronel F., A Practical Building Block for Solving Agreement Problems in Asynchronous Distributed Systems. *Proc. 17th IEEE Int. Performance, Computing and Communications Conference*, Phoenix (AZ), pp. 25-31, February 1998.

[8] Kihlstrom K.P., Moser L.E. and Melliar-Smith P.M., Solving Consensus in a Byzantine Environment Using an Unreliable Fault Detector. *Proc. First Int. Symposium on Principles of Distributed Systems (OPODIS'97)*, Hermes Ed., Chantilly (France), pp. 61-76, December 1997.

[9] Malkhi D. and Reiter M., Unreliable Intrusion Detection in Distributed Computations. In *Proc. of the 10th IEEE Computer Security Foundations Workshop*, pp. 116–124, Rockport (MA), June 1997.

[10] Pease L., Shostak R. and Lamport L., Reaching Agreement in Presence of Faults. *Journal of the ACM*, 27(2):228-234, 1980.

[11] Schiper A., Early Consensus in an Asynchronous System with a Weak Failure Detector. *Distributed Computing*, 10:149-157, 1997.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399