t-résilient snapshot immédiat
Carole Delporte-Gallet, Hugues Fauconnier, Sergio Rajsbaum, Michel Raynal

To cite this version:
Carole Delporte-Gallet, Hugues Fauconnier, Sergio Rajsbaum, Michel Raynal. t-résilient snapshot immédiat. ALGOTEL 2017 - 19èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications, May 2017, Quiberon, France. hal-01517320

HAL Id: hal-01517320
https://hal.archives-ouvertes.fr/hal-01517320
Submitted on 3 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Dans un système de \(n\) processus communiquant par mémoire partagée, un snapshot immédiat est un état de la mémoire assurant que si l’écriture réalisée par \(q\) est dans le snapshot de \(p\) alors le snapshot de \(p\) contient le snapshot obtenu par \(q\). Un snapshot immédiat peut être réalisé dans un système où au plus \(n-1\) processus peuvent tomber en panne. Dans un système où \(t\) processus peuvent tomber en panne, un processus peut obtenir un snapshot contenant les valeurs écrites par jusqu’à \(n-t\) processus. On définit ainsi le \(t\)-résilient snapshot immédiat et on établit des liens entre ce problème et les problèmes d’accord comme le consensus et le \(k\)-accord.

Mots-clés : Mémoire partagé, snapshot immédiat, tolérance aux pannes, consensus, \(k\)-accord.

1 Introduction

Nous considérons un modèle de calcul distribué de \(n\) processus asynchrones qui communique par le biais de la mémoire partagée. Un snapshot immédiat est une notion qui a été introduite il y a une dizaine d’années [1, 2] et a été intensivement étudiée par la communauté des algorithmes distribués. Un snapshot immédiat est une structure de données initialement vide qui peut contenir jusqu’à \(n\) pairs (un pair par processus), chaque pair composé d’un index de processus et d’une valeur. Cette structure apporte aux processus deux opérations, denotées update et snap. L’invocation \(\text{update}(v)\) par un processus \(p\) ajoute le pair \((i, v)\) à la structure et l’opération \(\text{snap}\) retourne tous les pairs écrits dans la structure.

Les \(t\)-résilients snapshots immédiats ont été introduits dans [3, 11], et plus tard dans [5, 15]. Il s’agit d’une variante du snapshot immédiat. Un invokation \(\text{write}\snapshot\) par un processus \(p\) ajoute le pair \((i, v)\) à la structure et retourne un ensemble de pairs \(\text{view}\) appartenant à la structure de telle sorte que si \((j, w)\) est dans \(\text{view}\), alors \(\text{view}\subseteq\text{view}\) et le pair \((j, w)\) est dans \(\text{view}\).

L’originalité de ce modèle d’itération des snapshots immédiats est la suivante. Il a été montré par Borowsky et Gafni dans [5], que ce modèle est équivalent à l’habitude de lecture/écriture wait-free model ((\(n-1\))-crash model) pour la résolvabilité. Il utilise le fait que ses tours sont plus structurés et plus faciles à analyser que les tours du modèle basique read/write shared memory model [14]. Il est aussi le cadre du mode combinatoire topologie approach pour le calcul distribué (e.g., [10]). Par conséquent, les IS objets constituent le fondement algorithmique du modèle de calcul itératif distribué.

Quand considérant le \(t\)-crash \(n\)-process model where \(t < n\), and assuming that each correct process writes a value, a process may wait for values written by \((n-t)\) processes without risking being blocked forever.

1 This paper is an extended abstract of [7]; the original title is \(t\)-résilient snapshot immédiat.

2 Carole Delporte-Gallet, Hugues Fauconnier and Michel Raynal are supported by ANR DESCARTES
(while it would return a set of \( x \) pairs with \( 1 \leq x \leq n \) if the object was an immediate snapshot object). Hence, a \( t \)-immediate snapshot object allows processes to obtain as much information as possible from the other processes while guaranteeing progress.

The obvious question is then the implementability of a \( t \)-immediate snapshot object in the \( \tau \)-crash \( n \)-process model.

Implementations of an \((n - 1)\)-immediate snapshot object is described in [3]. For the other values of \( \tau \) (\( 0 < \tau < n - 1 \)), this question is answered in this paper, which shows that it is impossible to implement a \( t \)-IS object in a \( \tau \)-crash \( n \)-process model when \( 0 < \tau < n - 1 \). More precisely we prove that implementing a \( t \)-IS object is equivalent\(^5\) to implementing consensus when \( \tau < n/2 \) and enables to implement \((2\tau - n + 2)\)-Set agreement when \( \tau/2 \leq \tau < n - 1 \).

At first glance, this impossibility result may seem surprising. First, an IS object is a snapshot object (a) whose operations update and snap are glued together in a single operation \texttt{write\_snapshot()}\(^1\), and (b) satisfying an additional property linking the sets of pairs returned by concurrent invocations. Then, as already indicated, a \( \tau \)-IS object is an IS object such that the sets returned by \texttt{write\_snapshot()} contain at least \((n - \tau)\) pairs. This property on the sets returned by a snapshot object can be trivially implemented in a \( \tau \)-crash \( n \)-process model. Hence, while a \( \tau \)-snapshot object can be implemented in the \( \tau \)-crash \( n \)-process model, a \( \tau \)-IS object cannot when \( 0 < \tau < n - 1 \). Second, because in general smallest is the number of possible faults easiest is the implementation. But in the \( \tau \)-IS, a small number of failure, induces a harder problem each output set must have the size at least \( n - \tau \).

2 Model

We consider a distributed computing model of \( n \geq 3 \) asynchronous sequential processes denoted \( p_1, ..., p_n \) among which any subset of up to \( \tau \) (\( 0 < \tau < n \)) processes may crash. The processes cooperate by reading and writing Single-Writer Multi-Reader atomic read/write registers [13].

One-shot immediate snapshot object. An immediate snapshot object (IS) is a set, initially empty, that will contain pairs made up of a process index and a value. Let us consider a process \( p_i \) that invokes \texttt{write\_snapshot(\textit{v})}\. This invocation adds the pair \((i, \textit{v})\) to the object, and returns to \( p_i \) a set, called view and denoted \texttt{view}\(i\), such that the sets returned to the processes collectively satisfy the following properties.

- Termination. The invocation of \texttt{write\_snapshot()} by a correct process terminates.
- Self-inclusion. \( \forall \ i: \ (\langle i, v \rangle \in \texttt{view}_i \) \)
- Validity. \( \forall \ i: \ (\langle j, v \rangle \in \texttt{view}_i \) \( \Rightarrow \ p_j \) invoked \texttt{write\_snapshot(\textit{v})}\.\)
- Containment. \( \forall \ i, j: \ \texttt{view}_i \subseteq \texttt{view}_j \) \( \lor \ \texttt{view}_j \subseteq \texttt{view}_i \).\)
- Immediacy. \( \forall \ i, j: \ (\langle i, v \rangle \in \texttt{view}_j \) \( \Rightarrow \ \texttt{view}_i \subseteq \texttt{view}_j \).

\( k \)-Set agreement. \( k \)-Set agreement was introduced by S. Chaudhuri [6], it generalizes consensus which corresponds to the case \( k = 1 \). A \( k \)-Set agreement object is a one-shot object that provides the processes with a single operation \texttt{propose\_k()}\. This operation allows the invoking process \( p_i \) to propose a value it passes as an input parameter (called \texttt{proposed\_value})\(^2\), and obtain a value (called \texttt{decided\_value})\(^3\). The object is defined by the following set of properties.

- Termination. The invocation of \texttt{propose\_k()} by a correct process terminates.
- Validity. A decided value is a proposed value.
- Agreement. No more than \( k \) different values are decided.

It is shown in [8, 4, 11, 16] that the problem is impossible to solve if \( k \leq \tau \).

\( t \)-Immediate Snapshot. A \( t \)-immediate snapshot object (denoted by \( t \)-IS) is an immediate snapshot object with the following additional property.

- Output size. The set \texttt{view} obtained by a process is such that \( |\texttt{view}| \geq n - t \).

3 Results

\( t \)-Immediate Snapshot is Impossible if \( 0 < \tau < n - 1 \). This section presents an algorithm Figure 1 to achieve (1) consensus from \( t \)-IS for \( 0 < \tau < n/2 \), and (2) \( k \)-Set agreement (in short \( k \)-SA) from \( t \)-IS for

\( ^5 \) A is equivalent to B if A can be (computationally) reduced to B and reciprocally.
t-resilient immediate snapshot

\begin{algorithm}
\begin{algorithmic}[1]
\State \textbf{PROPOSE}(v)
\begin{algorithmic}
\State \textbf{begin}
\State \texttt{view} $\leftarrow$ \texttt{I.write\_snapshot}(v); /* I shared immediate snapshot*/
\State \texttt{VIEW}[i] $\leftarrow$ \texttt{view}; /* VIEW is a shared array*/
\State \texttt{wait}(\{|j| \text{ such that } \texttt{VIEW}[j] \neq \bot\}| = t + 1);
\State \texttt{let view} be the smallest of the previous \((t + 1)\) views;
\State \texttt{return} (smallest proposed value in view)
\State \textbf{end PROPOSE}
\end{algorithmic}
\end{algorithmic}
\caption{Solving consensus from t-IS if 0 < t < n/2(code for p1)}
\end{algorithm}

\begin{algorithm}
\begin{algorithmic}[1]
\State \textbf{write\_snapshot}(v_i)
\begin{algorithmic}
\State \textbf{begin}
\State \texttt{S.update}(i, v_i); /* S shared snapshot*/
\State \texttt{view} $\leftarrow$ \texttt{\theta}; \texttt{dec} $\leftarrow$ \texttt{\theta}; \texttt{k} $\leftarrow$ \texttt{-1}; launch the tasks T1 and T2.
\State \textbf{Task T1}:
\begin{algorithmic}
\State \textbf{repeat} \texttt{k} $\leftarrow$ \texttt{k + 1}
\State \textbf{do} \texttt{aux} $\leftarrow$ \texttt{S.snapshot()}\textbf{ until} (\texttt{dec} $\subseteq$ \texttt{aux} $\land$ \texttt{|aux|} $\geq$ n $- t$)
\State \texttt{dec} $\leftarrow$ \texttt{CONS}[\texttt{k}], \texttt{propose}_1(\texttt{aux})
\State \textbf{if} \left(\langle i, v_i \rangle \in \texttt{dec} \land (\texttt{view} = \texttt{\theta})\right) \texttt{then view} $\leftarrow$ \texttt{dec} \textbf{end if}
\State \textbf{until} \texttt{|aux|} $=$ n
\State \textbf{end task T1}
\end{algorithmic}
\State \textbf{Task T2} : \texttt{wait} (\texttt{view} $\neq$ \texttt{\theta}) ; \texttt{return} (\texttt{view}) \textbf{end task T2}.
\State \textbf{end write\_snapshot}
\end{algorithmic}
\caption{Implementing t-IS with consensus (code for p0)}
\end{algorithm}

\(k = 2t - n + 2\) (e.g., \((n - 2)\)-SA agreement from \((n - 2)\)-IS if \(t = n - 2\)). As these problems are impossible to solve [4, 11, 16], we show that it is impossible to implement a t-IS object when 0 < t < n - 1.

Intuitively this algorithm works because there is a set of at least \(\ell \geq n - t\) processes, that obtained the same view \texttt{min\_view} (or crashed before returning from \texttt{write\_snapshot}()), and this view is the smallest view obtained by a process and its size is |\texttt{min\_view}| = \(\ell\). If 0 < t < n/2, as \(\ell \geq n - t\) and \((n - t) + (t + 1) > n\), it follows from the waiting predicate, that, any process that executes line 5, obtains a copy of \texttt{min\_view}, and consequently we have \texttt{view} = \texttt{min\_view} at line 5. It follows that no two processes can decide different values. If \(n/2 \leq t < n - 1\), we have \(n - t \leq t\). The \(m = (n - t) - 1\) biggest views will never be selected by the processes, and consequently these processes obtain at most \(t - m = t - ((n - t) - 1) = 2t - n + 1\) different smallest views. Hence, these processes may decide at most \(2t - n + 1\) different values.  

\textbf{Theorem 1} A t-IS object cannot be implemented if 0 < t < n - 1.

\textbf{From Consensus to t-IS} if 0 < t < n - 1. While a snapshot object is atomic (operations on a snapshot object can be linearized [12]), an IS object and a \(k\)-immediate snapshot objects are not atomic (its operations cannot always be linearized).

So we cannot apply the universality result of Herlihy [9], and we have to write a specific algorithm given Figure 2.

\textbf{t-Immediate Snapshot and \(k\)-Set agreement.} With the algorithms Figures 1 and 2, we get:

\textbf{Theorem 2} Consensus and t-IS are equivalent if 0 < t < n/2.

We have shown that from t-IS when \(n/2 \leq t < n - 2\) we can implement (2t - n + 2)-Set agreement. Can we do consensus as in the case 0 < t < n/2? By a simulation argument, we show that consensus is not solvable with t-immediate snapshot when \(n/2 \leq t < n\) proving that the computational power of t-immediate snapshot when 0 < t < n/2 is strictly stronger than the computational power of t-immediate snapshot when n/2 \leq t < n.
Theorem 3 If $0 < t < n/2$ then $t$-IS can implement $(2t - n + 2)$-Set agreement and cannot implement consensus. Consensus implements $t$-IS.

<table>
<thead>
<tr>
<th>$1 \leq t &lt; n/2$</th>
<th>$n/2 \leq t &lt; n - 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$-IS and consensus are equivalent</td>
<td>$t$-IS implements $(2t - n + 2)$-Set agreement</td>
</tr>
<tr>
<td>$t$-IS does not implement consensus</td>
<td>consensus implements $t$-IS</td>
</tr>
</tbody>
</table>

Table 1: Summary of results presented in the paper

4 Conclusion

The paper has shown that, while it is possible to build an $(n - 1)$-IS object in the asynchronous read/write $(n - 1)$-crash model, it is impossible to build a $t$-IS object in an asynchronous read/write $t$-crash model when $0 < t < n - 1$. It follows that the notion of an Iterative immediate snapshot distributed model seems inoperative for these values of $t$. The results of the paper are summarized in Table 1.

Interestingly, this study shows that there are two contrasting impossibility results in asynchronous read/write $t$-crash $n$-process systems. Consensus is impossible as soon as $t > 0$, while $t$-immediate snapshot is impossible as soon as $t < n - 1$.

As a final remark, some computability problems remain open. As an example, is it possible to implement a $t$-IS object from $(2t - n + 2)$-Set agreement?

Références


