Allocation adaptative de registres en utilisant un nombre linéaire de registres
Carole Delporte-Gallet, Hugues Fauconnier, Eli Gafni, Leslie Lamport

To cite this version:

HAL Id: hal-00978860
https://hal.archives-ouvertes.fr/hal-00978860
Submitted on 19 Jun 2014

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.
Allocation adaptative de registres en utilisant un nombre linéaire de registres†

Carole Delporte-Gallet¹‡ and Hugues Fauconnier¹§ and Eli Gafni³ and Leslie Lamport⁴

¹ LIAFA, U. Paris Diderot, France
² Computer Science Department, UCLA, USA
³ Microsoft Research

On présente un algorithme adaptatif dans lequel les processus utilisent des registres multi-écrivains multi-lecteurs. Cet algorithme permet à chaque processus d’obtenir un accès exclusif à un registre dont il sera l’écrivain unique et que tous les processus pourront lire. L’algorithme est adaptatif : il ne connait pas a priori le nombre de processus qui vont demander un accès exclusif en écriture à un registre. C’est le premier algorithme permettant d’obtenir ce résultat en utilisant des registres dont le nombre est une fonction linéaire du nombre de participants. Les précédents algorithmes adaptatifs utilisent au moins $\Theta(n^{3/2})$ registres.

Keywords: shared memory, read/write registers, distributed algorithms, wait-free, space complexity, renaming.

1 Introduction

One way to implement multiprocess synchronization is by providing each process with a single-writer, multi-reader atomic register (SWMR) that it can write and other processes can read. We present an adaptive algorithm to implement such a system of registers with an array of multi-writer multi-reader atomic (MWMR) registers whose length is linear in the number of participating processes. The algorithm is non-blocking unless an unbounded number of processes initiate operations.

An adaptive algorithm, also called a uniform algorithm [Gaf02], is one that does not know the number of potentially participating processes. Equivalently, it is an algorithm whose cost is a function not of the total number of processes but of the number of processes that actually participate in the algorithm. For the SWMR registers, this is the number of processes that actually perform a read or write operation. Our goal is to minimize the number of MWMR registers, and our algorithm uses a number that is linear in the number of participants. No a priori bound on this number is assumed.

Why do we find this algorithm interesting? There are simpler algorithms that assume stronger communication primitives—for example, test and set registers—but MWMR registers are the weakest ones for which we know that an adaptive algorithm is possible. More efficient randomized algorithms are possible, but our algorithm is always correct, not just correct with high probability. There is a trivial way to implement a collection of SWMR registers with an array $C$ of MWMR registers. The $i^{th}$ process simply uses $C[i]$ as its register. Of course, this algorithm uses an unbounded number of registers. The obvious way to make the number of registers linear in the number of participating processes is by having the processes first execute an adaptive renaming algorithm [ABND + 90, BG93] in which each participating process is assigned a unique number from 0 to $M$ for some $M$ that depends linearly on the number of participants. A process assigned the number $j$ then uses $C[j]$ as its register. However, we know of few renaming algorithms that do not assume a collection of SWMR registers already allocated to processes [Asp10, AF02, MA95]. Those algorithms are all based on the grid-network of “splitters” proposed by Anderson and Moir [MA95]. Of these,

†This work has been accepted for publication at DISC2013 [DGFGL13].
‡Supported by ANR DISPLEXITY.
§Supported by ANR DISPLEXITY.
the more space-efficient is an improvement of Aspnes [Asp10] that requires $\Theta(k^{3/2})$ MWMR registers for $k$ participating processes. Even though the renaming algorithm is used only to determine the assignment of processes to elements of the array $C$, the values in those $\Theta(k^{3/2})$ registers must be maintained forever because additional processes may enter the system at any time. (Reclaiming the space requires knowing an a priori bound on the number of processes that might participate.) Thus, our algorithm is the first that implements a collection of SWMR registers with $O(k)$ MWMR registers.

Almost all previous methods for making an algorithm adaptive start by using one of several renaming algorithms [AAD+93, AST99, And94, ABND+90, BG93]. It has generally been assumed that this is the only way to implement an adaptive algorithm [AW98]. Based on an idea in [DGFGR13], our implementation avoids the use of a renaming algorithm to begin reliable communication. Instead, participating processes first announce their presence by using a non-blocking one-shot limited-snapshot algorithm that we call the GFX (Generalized Fast eXclusion) protocol, which can be viewed as generalizing [Lam87] from 1-concurrency to $k$-concurrency. The snapshot is limited to having the property that two snapshots of the same size coincide. It need not ensure that snapshots of different sizes are related by containment. To perform a read or write operation to a register, a process first reads the posted snapshots to find the number $n$ of participants that have announced their presence, and it executes an algorithm [DGFRG13] that assumes at most $n$ processes. It then reads the number of participants again, finishing the operation if that number still equals $n$. Otherwise, the process repeats the $n$-process algorithm for the new value of $n$. While we use this approach to implement renaming, it can be used to provide an adaptive implementation of any task.

By using our adaptive algorithm for implementing a collection of SWMR registers, we can solve any task under the assumption of finite arrival [GMT01]. In particular, using existing algorithms, we can implement adaptive renaming with a linear range [ABND+90, BG93]. This in turn allows us to allocate unique registers to processes with a number of registers linear in the number of participants. With register allocation, we can implement a collection of SWMR registers with wait-free read and write operations rather than just non-blocking ones. For many tasks of high read-write complexity, doing renaming first may reduce the step complexity of an adaptive algorithm.

We ignore time complexity—the number of steps taken by the algorithm. Our algorithm is executed just once, to assign SWMR registers to processes; it adds nothing to the cost of using those registers. Since space used by an adaptive algorithm cannot be reclaimed, it is perhaps more important than time complexity. Still, optimal time complexity is an interesting problem that remains unsolved.

In the non-adaptive case, it has been shown that at least $n$ registers are required to implement $n$ SWMR registers [DGFGR13], so the linear number of registers used by our algorithms is optimal up to a constant factor. We originally believed that adaptive algorithms required more than a linear number of registers, and we tried to derive such a lower bound on the number of registers, independent of their size. When the difficulty is caused by processes stepping on each other because of the lack of a priori coordination, size of the registers is not a factor. (See the lower bound for consensus [FHS98].) We were therefore surprised to discover our algorithm.

We precisely describe our algorithms in the PlusCal algorithm language [Lam09]. A PlusCal expression can be any TLA + formula [Lam02], and a PlusCal algorithm is automatically translated to a TLA + specification that defines the algorithm’s formal meaning.

We have written formal, mechanically-checked TLA + correctness proofs of the safety properties of the
2 Algorithms

---algorithm SnapShot
{ variables result = \{ p ∈ Proc \rightarrow \{ \} \}, A2 = \{ i ∈ Nat \rightarrow \{ \} \}, A3 = \{ i ∈ Nat \rightarrow \{ \} \} ;
process ( Pr ∈ Proc )
variables myVals = \{ \}, known = \{ \}, notKnown = \{ \},
lnbpart = 0, npart = 0, nextout = \{ \}, out = \{ \} ;
{ a : with \{ P ∈ \{ Q ∈ \text{SUBSETProc} : ⊓ self ∈ Q ∧ ∀ p ∈ Proc \setminus \{ self \} : ∨ \text{Cardinality}(result[p]) ≠ \text{Cardinality}(Q) \vee Q = result[p] \} ;
    \{ result[\{ self \}] := P ; \}
} A2(\text{Cardinality}(result[\{ self \}]) - 1) := result[\{ self \}] ;

\begin{align*}
b & : \text{while ( TRUE )} \quad \{ \text{with (} v \in \text{Val}) \{ \text{myVals := myVals} \cup \{ v \} \} ; \\
  & \text{known := myVals} \cup \text{known} ; \\
  & \text{npart} := \text{Cardinality}(\text{NUnion}(A2)) ;
\end{align*}

\begin{align*}
c & : \text{lnbpart} := \text{npart} ; \\
  & \text{notKnown} := \{ i ∈ 0 \ldots (\text{npart} - 1) : \text{known} ≠ A3[i] \} ; \\
  & \text{if (notKnown} ≠ \{ \}) \quad \{ d : \text{with (} i ∈ \text{notKnown} \} \\
  & \quad \{ A3[i] := \text{known} \} ;
\end{align*}

goto c \\
\text{e} : \text{npart} := \text{Cardinality}(\text{NUnion}(A2)) ;
\text{if (lnbpart} = \text{npart} \} \{ \text{out} := \text{known} \} \quad \text{/*returned value */}
\text{else (goto c)}
}

\text{Figure 2: Algorithm SnapShot.}

A sequence of SWMR registers is implemented using an algorithm we call SnapShot. This algorithm begins with Algorithm GFX that we describe below.

\text{Algorithm GFX}

Algorithm GFX, described in Figure 1, solves the following weaker version of the snapshot task [AAD’93] :
A process \( p \) that executes the algorithm must return a set \( F_p \) of participants such that
- \( F_p = F_q \) for any \( p \),
- \( |F_p| = |F_q| \) implies \( F_p = F_q \) for any \( p \) and \( q \), where \( |F| \) is the cardinality of the set \( F \).

The variables \text{known} and \text{notKnown} are local to \text{self} (the current process) and cannot be read or written by other processes. Variable \text{known} stores the set of processes known to \text{self} process, and \text{notKnown} stores a set of array indices (natural numbers). The values of these process-local variables are arrays indexed by the set \text{Proc}. The other new notations used in this algorithm are : \text{Nat} is the set of natural numbers, \( i \ldots j \) is the set of integers \( k \) with \( i ≤ k ≤ j \), the statement \text{with (} x \in S \{ \Sigma \} \text{executes} \Sigma \) with an arbitrary element of \( S \) substituted for \( x \) and the operator \text{NUnion} is defined by \( \text{NUnion}(A) ≝ \cup \text{Union}\{A[i] : i ∈ \text{Nat}\} \).

Evaluation of that expression is implemented by atomically reading the array \( A \). Observe that although \text{result} is a global variable, \text{result}[p] is accessed only by process \( p \).

\text{Algorithm SnapShot}

The SnapShot algorithm maintains a set \( S \) of values that is initially empty. It provides a \text{snap} operation whose argument is a value \( v \). Executing \text{snap}(v) atomically adds \( v \) to \( S \) and returns the current value of \( S \). This allow to simulate for each process a SWMR register.

Let’s suppose that there is a \text{count} operation that a process \( p \) can call to learn the number of participants that can be executing a \text{snap} operation. To perform a \text{snap} operation, a process \( p \) first executes \text{count} to
obtain a bound \( n \) on the number of participants. It then writes in the first \( n \) registers of \( A_3 \). If a read of \( A_3 \) obtains a value \( F \) such that \( A_3[0] = \cdots = A_3[n - 1] = F \), process \( p \) executes the count operation again. If that execution returns the same number \( n \) of participants, then the snap operation completes and returns the value \( F \). Otherwise, the process continues the procedure, replacing \( n \) with the new value returned by count.

We still have to implement the count operation. We do that by using algorithm GIFX and a second array \( A_2 \) of registers. When a participant \( p \) arrives, before performing any snap operation it (i) executes GIFX to obtain a set \( S \) of participants, which includes itself, and (ii) writes (the processes in) \( S \) in \( A_2[|S| - 1] \). The correctness property of GIFX implies that no other value can ever be written in \( A_2[|S| - 1] \). Since the processes written in \( A_2 \) are all participants and every participant is written in \( A_2 \), the set of all processes in \( A_2 \) includes all participants that can write to \( A_3 \). The count operation is then performed by reading \( A_2 \) and counting the number of (distinct) processes read.

Algorithm SnapShot appears in Figure 2. We have represented the code of GIFX in SnapShot by the corresponding code of its specification in TLA+.

Références


