CHAPTER 5. SHARED MEMORY

5.1 Introduction

In distributed computing, various distributed systems exist. So, for the focus of the course we will be working on loosely-coupled distributed systems such as the Internet, where nodes asynchronously communicate by exchanging messages. The "opposite" model is a tightly-coupled parallel computer where nodes access a common memory totally synchronously. In a distributed system, the shared memory is managed by the operating system (e.g., a Distributed Random Access Machine (PRAM)).

A shared memory system consists of asynchronous processes that access a common (shared) memory. A process can atomically access a register in the shared memory through a set of predefined operations. An atomic modification appears to the rest of the system instantaneously. Apart from this shared memory, processes can also have some local (private) memory.

Definition 5.1 (Shared Memory) A shared memory system is a system that consists of asynchronous processes that access a common (shared) memory. A process can atomically access a register in the shared memory through a set of predefined operations. An atomic modification appears to the rest of the system instantaneously. Apart from this shared memory, processes can also have some local (private) memory.

Remarks:

- Various shared memory systems exist. A main difference is how they allow processes to access the shared memory. All systems can atomically read or write a shared register.
- Most systems do allow advanced read-modify-write (RMW) operations, for example:
  - test-and-set($R$): $t := R; R := 1; \text{return } t$
  - fetch-and-add($R$, x): $t := R; R := R + x; \text{return } t$
  - compare-and-swap($R$, x, y): if $R = x$ then $R := y; \text{return true}$; else $\text{return false}$.

5.2 Mutual Exclusion

A classic problem in shared memory systems is mutual exclusion. We are given a number of processes, each of which occasionally needs to access the same resource. The resource may be a shared variable or a more general object such as a data structure or a shared printer. The catch is that only one process at a time is allowed to access the resource. More formally,

Definition 5.2 (Mutual Exclusion) We are given a number of processes, each with a shared variable.

A mutual exclusion algorithm must ensure that only one process at a time can access the shared variable, preventing race conditions.

The following holds...
5.2. MUTUAL EXCLUSION

- Mutual Exclusion: At all times at most one process is in the critical section.
- No deadlock: If some process manages to get to the entry section, later some (possibly different) process will get to the critical section.

Sometimes we in addition ask for
- No lockout: If some process manages to get to the entry section, later some (possibly different) process will get to the critical section.
- Unobstructed exit: No process can get stuck in the exit section.

Using RMW primitives one can build mutual exclusion algorithms quite easily. Algorithm 23 shows an example with the test-and-set primitive.

Algorithm 23 Mutual Exclusion: Test-and-Set

Input: Shared register \( R = 0 \)

\(<\text{Entry}>\)
1. repeat
2. \( r := \text{test-and-set}(R) \)
3. until \( r = 0 \)
\(<\text{Critical Section}>>\)
4. \( R = 0 \)
\(<\text{Exit}>\)
5. \( R = 0 \)
\(<\text{Remainder Code}>\)
6. . .

Theorem 5.3. Algorithm 23 solves the mutual exclusion problem as in Definition 5.2.

Proof. Mutual exclusion follows directly from the test-and-set definition: Initially \( R = 0 \). Let \( p_i \) be the \( i \)th process to successfully execute the test-and-set, where successfully means that the result of the test-and-set is 0. This happens at time \( t_i \). At time \( t_i \) process \( p_i \) resets the shared register \( R \) to 0. Between \( t_i \) and \( t'_i \) no other process can successfully test-and-set, hence no other process can enter the critical section concurrently.

Proving no deadlock works similar. One of the processes loitering in the entry section will successfully test-and-set as soon as the process in the critical section exited. Since the exit section only consists of a single instruction (no potential infinite loops) we have unobstructed exit.

Remarks:
- No lockout, on the other hand, is not given by this algorithm. Even with only two processes there are asynchronous executions where always the same process wins the test-and-set.
- Algorithm 23 can be adapted to guarantee fairness (no lockout), essentially by ordering the processes in the entry section in a queue.

Algorithm 24 Mutual Exclusion: Peterson’s Algorithm

Initialization: Shared registers \( W_0, W_1, II \), all initially 0.

Code for process \( p_i \), \( i \in \{0, 1\} \)

\(<\text{Entry}>\)
1. \( W_i := 1 \)
2. \( II := 1 - i \)
3. repeat until \( II = i \) or \( W_{1-i} = 0 \)
\(<\text{Critical Section}>>\)
4. . .
\(<\text{Exit}>>\)
5. \( W_i := 0 \)
\(<\text{Remainder Code}>>\)
6. . .

Remarks:
- A natural question is whether one can achieve mutual exclusion with only reads and writes, that is without advanced RMW operations. The answer is yes!

Our read/write mutual exclusion algorithm is for two processes \( p_0 \) and \( p_1 \) only. In the remarks we discuss how it can be extended. The general idea is that process \( p_i \) has to mark its desire to enter the critical section in a “want” register \( W_0 \) by setting \( W_i = 1 \). Only if the other process is not interested \( (W_{1-i} = 0) \) access is granted. This however is too simple since we may run into a deadlock. This deadlock (and at the same time also lockout) is resolved by adding a priority variable \( II \). See Algorithm 24.

Algorithm 24 Mutual Exclusion: Peterson’s Algorithm

\(<\text{Entry}>>\)
1. \( W_i := 1 \)
2. \( II := 1 - i \)
3. repeat until \( II = i \) or \( W_{1-i} = 0 \)
\(<\text{Critical Section}>>\)
4. . .
\(<\text{Exit}>>\)
5. \( W_i := 0 \)
\(<\text{Remainder Code}>>\)
6. . .

Theorem 5.4. Algorithm 24 solves the mutual exclusion problem as in Definition 5.2.

Proof. The shared variable \( II \) elegantly grants priority to the process that passes line 2 first. If both processes are competing, only process \( p_0 \) can access the critical section because of \( II \). The other process \( p_{1-i} \) cannot access the critical section because \( W_{1-i} = 1 \) (and \( II \neq 1 - II \)). The only other reason to access the critical section is because the other process is in the remainder code (that is, not interested). This proves mutual exclusion!

No deadlock comes directly with \( II \). Process \( p_0 \) gets direct access to the critical section, no matter what the other process does.

Since the exit section only consists of a single instruction (no potential infinite loops) we have unobstructed exit.

Thanks to the shared variable \( II \) also no lockout (fairness) is achieved: If a process \( p_i \) looses against its competitor \( p_{1-i} \) in line 2, it will have to wait until the competitor resets \( W_{1-i} = 0 \) in the exit section. If process \( p_i \) is unlucky it will not check \( W_{1-i} = 0 \) early enough before process \( p_{1-i} \) sets \( W_{1-i} = 1 \) again in line 1. However, as soon as \( p_{1-i} \) hits line 2, process \( p_i \) gets the priority due to \( II \), and can enter the critical section.
5.3 Store & Collect

5.3.1 Problem Definition

In this section, we will look at a second shared memory problem that has an elegant solution. Informally, the problem can be stated as follows. There are \( n \) processes \( p_1, \ldots, p_n \). Every process \( p_i \) has a read/write register \( R_i \) in the shared memory where it can store some information that is destined for the other processes. Further, there is an operation by which a process can collect (i.e., read) the values of all the processes that stored some value in their register.

We say that an operation \( op_1 \) precedes an operation \( op_2 \) if \( op_1 \) terminates before \( op_2 \) starts. An operation \( op_2 \) follows an operation \( op_1 \) if \( op_1 \) precedes \( op_2 \).

Definition 5.5 (Collect). There are two operations: \( \text{store} \) and \( \text{collect} \). A \( \text{store} \) operation returns the latest value of its register \( R_i \). A \( \text{collect} \) operation returns a view, a partial function \( V \) from the set of processes to a set of values, where \( V(p_i) \) is the latest value stored by \( p_i \), for each process \( p_i \). For a \( \text{collect} \) operation \( \text{cop} \), the following validity properties must hold for every process \( p_i \):

- If \( V(p_i) = \bot \), then no \( \text{store} \) operation by \( p_i \) precedes \( \text{cop} \).
- If \( V(p_i) = v \neq \bot \), then \( v \) is the value of a \( \text{store} \) operation \( \text{sop} \) by \( p_i \) that does not follow \( \text{cop} \), and there is no \( \text{store} \) operation by \( p_i \) that follows \( \text{sop} \) and precedes \( \text{cop} \).

Hence, a \( \text{collect} \) operation \( \text{cop} \) should not read from the future or miss a preceding \( \text{store} \) operation \( \text{sop} \).

We assume that the read/write register \( R_i \) of every process \( p_i \) is initialized to \( \bot \). We define the step complexity of an operation \( op \) to be the number of accesses to registers in the shared memory. There is a trivial solution to the collect problem as shown by Algorithm 25.

Algorithm 25 Collect: Simple (Non-Adaptive) Solution

\begin{verbatim}
Operation \text{store}(val) (by process \( p_i \)):
1. \( R_i := \text{val} \)

Operation \text{collect}:
2. for \( i = 1 \) to \( n \) do
3. \( V(p_i) := R_i \) // read register \( R_i \)
4. end for
\end{verbatim}

Remarks:
- Algorithm 25 clearly works. The step complexity of every \( \text{store} \) operation is 1, the step complexity of a \( \text{collect} \) operation is \( n \).
- At first sight, the step complexities of Algorithm 25 seem optimal. Because there are \( n \) processes, there are scenarios in which a \( \text{collect} \) operation needs to read all \( n \) registers. However, there are also scenarios in which the step complexity of the \( \text{collect} \) operation seems very costly. Assume that there are only two processes \( p_1 \) and \( p_2 \) that have stored a value in their registers \( R_1 \) and \( R_2 \). In this case, a \( \text{collect} \) is performed on both registers, so the \( \text{collect} \) operation seems very costly. Assume that there are only two processes \( p_1 \) and \( p_2 \) that have stored a value in their registers \( R_1 \) and \( R_2 \). In this case, a \( \text{collect} \) is performed on both registers, so the \( \text{collect} \) operation seems very costly.
- Assume that up to a certain time \( k \), \( k \leq n \) processes have finished or started at least one operation. We call an operation \( op \) at time \( k \) adaptive to contention if the step complexity of \( op \) only depends on \( k \) and is independent of \( n \).
- In the following, we will see how to implement adaptive versions of \( \text{store} \) and \( \text{collect} \).

5.3.2 Splitters

Algorithm 26 Splitter Code

\begin{verbatim}
Shared Registers: \( X : \{\bot\} \cup \{1, \ldots, n\} \); \( Y \) : boolean
Initialization: \( X := \bot \); \( Y := \text{false} \)

Splitter access by process \( p_i \):
1. if \( Y \) then
2. \( X := i \)
3. \( Y := \text{true} \)
4. return right
5. else
6. \( Y := \text{false} \)
7. return stop
8. if \( X = i \) then
9. \( Y := \text{true} \)
10. return left
11. end if
12. end if
\end{verbatim}

To obtain adaptive collect algorithms, we need a synchronization primitive, called a splitter.

Definition 5.6 (Splitter). A splitter is a synchronization primitive with the following characteristic. A process entering a splitter exits with either \( \text{stop}, \text{left} \), or \( \text{right} \). If \( k \) processes enter a splitter, at most one process exits with \( \text{stop} \) and at most \( k - 1 \) processes exit with \( \text{left} \) and \( \text{right} \), respectively.

Hence, it is guaranteed that if a single process enters the splitter, then it obtains \( \text{stop} \), and if two or more processes enter the splitter, then there is at most one process obtaining \( \text{stop} \) and there are two processes that obtain
Algorithm 27 Adaptive Collect: Binary Tree Algorithm

Operation \( \text{store}(val) \) by process \( p_i \):
1. \( R_i := \text{val} \)
2. if first \( \text{store} \) operation by \( p_i \) then
3. \( v := \text{root node of binary tree} \)
4. \( \alpha := \text{result of entering splitter } S(v) \)
5. \( M_{S(v)} := \text{true} \)
6. while \( \alpha \neq \text{stop} \) do
7. if \( \alpha = \text{left} \) then
8. \( v := \text{left child of } v \)
9. else
10. \( v := \text{right child of } v \)
11. end if
12. \( \alpha := \text{result of entering splitter } S(v) \)
13. \( M_{S(v)} := \text{true} \)
14. end while
15. \( Z_{p_i} := 1 \)
16. end if

Operation \( \text{collect} \):
17. for all marked splitters \( S \) do
18. if \( Z_S \neq \text{false} \) then
19. \( i := Z_S; \ V(p_i) := R_i \) // read value of process \( p_i \)
20. end if
21. end for
22. // \( V(p_i) = \bot \) for all other processes

5.3. STORE & COLLECT

Figure 5.1: A Splitter

Algorithm 26 correctly implements a splitter.

Proof. Assume that \( k \) processes enter the splitter. Because the first process that checks whether \( Y = \text{true} \) in line 2 will find that \( Y = \text{false} \), not all processes return \( \text{right} \). Next, assume that \( i \) is the last process that sets \( X := i \). If \( i \) does not return \( \text{right} \), it will find \( X = i \) in line 6 and therefore return \( \text{stop} \). Hence, there is always a process that does not return \( \text{left} \). It remains to show that at most 1 process returns \( \text{stop} \). For the sake of contradiction, assume \( p_i \) and \( p_j \) are two processes that return \( \text{stop} \) and assume that \( p_i \) sets \( X := i \) before \( p_j \) sets \( X := j \). Both processes need to check whether \( Y = \text{true} \) before one of them sets \( Y = \text{true} \). Hence, they both complete the assignment in line 1 before the first one of them checks the value of \( X \) in line 6. Hence, by the time \( p_i \) arrives at line 6, \( X \neq i \) (\( p_j \) and maybe some other processes have overwritten \( X \) by them). Therefore, \( p_i \) does not return \( \text{stop} \) and we get a contradiction to the assumption that both \( p_i \) and \( p_j \) return \( \text{stop} \). □

5.3.3 Binary Splitter Tree

Assume that we are given \( 2^i - 1 \) splitters and that for every splitter \( S \), there is an additional shared variable \( Z_S : \{ \bot \} \cup \{ 1, \ldots, n \} \) that is initialized to \( \bot \) and an additional shared variable \( M_S : \{ \text{boolean} \} \) that is initialized to \( \text{false} \). We call a splitter \( S \) marked if \( M_S := \text{true} \). The \( 2^i - 1 \) splitters are arranged in a complete binary tree of height \( n - 1 \). Let \( S(v) \) be the splitter associated with a node \( v \) of the binary tree. The \( \text{store} \) and \( \text{collect} \) operations are given by Algorithm 27.

Theorem 5.8. Algorithm 27 correctly implements \( \text{store} \) and \( \text{collect} \). Let \( k \) be the number of participating processes. The step complexity of the first \( \text{store} \) of a process \( p_i \) is \( O(k) \), the step complexity of every additional \( \text{store} \) of \( p_i \) is \( O(1) \), and the step complexity of \( \text{collect} \) is \( O(k) \).

Proof. Because at most one process can stop at a splitter, it is sufficient to show that every process stops at some splitter at depth at most \( k - 1 \leq n - 1 \) when invoking the first \( \text{store} \) operation to prove correctness. We prove that at most \( k - 1 \) processes enter a subtree at depth \( i \) (i.e., a subtree where the root has distance \( i \) to the root of the whole tree). By definition of \( k \), the number of different values (i.e., either there is exactly one \text{stop} or there is at least one \text{left} and at least one \text{right}). For an illustration, see Figure 5.1. The code implementing a splitter is given by Algorithm 26.

Lemma 5.7. Algorithm 26 correctly implements a splitter.

Proof. Assume that \( k \) processes enter the splitter. Because the first process that checks whether \( Y = \text{true} \) in line 2 will find that \( Y = \text{false} \), not all processes return \( \text{right} \). Next, assume that \( i \) is the last process that sets \( X := i \). If \( i \) does not return \( \text{right} \), it will find \( X = i \) in line 6 and therefore return \( \text{stop} \). Hence, there is always a process that does not return \( \text{left} \). It remains to show that at most 1 process returns \( \text{stop} \). For the sake of contradiction, assume \( p_i \) and \( p_j \) are two processes that return \( \text{stop} \) and assume that \( p_i \) sets \( X := i \) before \( p_j \) sets \( X := j \). Both processes need to check whether \( Y = \text{true} \) before one of them sets \( Y = \text{true} \). Hence, they both complete the assignment in line 1 before the first one of them checks the value of \( X \) in line 6. Hence, by the time \( p_i \) arrives at line 6, \( X \neq i \) (\( p_j \) and maybe some other processes have overwritten \( X \) by them). Therefore, \( p_i \) does not return \( \text{stop} \) and we get a contradiction to the assumption that both \( p_i \) and \( p_j \) return \( \text{stop} \). □

5.3.3 Binary Splitter Tree

Assume that we are given \( 2^i - 1 \) splitters and that for every splitter \( S \), there is an additional shared variable \( Z_S : \{ \bot \} \cup \{ 1, \ldots, n \} \) that is initialized to \( \bot \) and an additional shared variable \( M_S : \{ \text{boolean} \} \) that is initialized to \( \text{false} \). We call a splitter \( S \) marked if \( M_S := \text{true} \). The \( 2^i - 1 \) splitters are arranged in a complete binary tree of height \( n - 1 \). Let \( S(v) \) be the splitter associated with a node \( v \) of the binary tree. The \( \text{store} \) and \( \text{collect} \) operations are given by Algorithm 27.

Theorem 5.8. Algorithm 27 correctly implements \( \text{store} \) and \( \text{collect} \). Let \( k \) be the number of participating processes. The step complexity of the first \( \text{store} \) of a process \( p_i \) is \( O(k) \), the step complexity of every additional \( \text{store} \) of \( p_i \) is \( O(1) \), and the step complexity of \( \text{collect} \) is \( O(k) \).

Proof. Because at most one process can stop at a splitter, it is sufficient to show that every process stops at some splitter at depth at most \( k - 1 \leq n - 1 \) when invoking the first \( \text{store} \) operation to prove correctness. We prove that at most \( k - 1 \) processes enter a subtree at depth \( i \) (i.e., a subtree where the root has distance \( i \) to the root of the whole tree). By definition of \( k \), the number of processes entering the splitter at depth 0 (i.e., at the root of the binary tree) is \( k \). For \( i > 0 \), the claim follows by induction because of the at most \( k - i \) processes entering the splitter at the root of a depth \( i \) subtree, at most \( k - i - 1 \) obtain \text{left} and \text{right}, respectively. Hence, at the latest when reaching depth \( k - 1 \), a process is the only process entering a splitter and thus obtains \( \text{stop} \). It thus also follows that the step complexity of the first invocation of \( \text{store} \) is \( O(k) \).

To show that the step complexity of \( \text{collect} \) is \( O(k) \), we first observe that the marked nodes of the binary tree are connected, and therefore can be traversed by only reading the variables \( Z_S \) associated to them and their neighbors. Hence, showing that at most \( 2k - 1 \) nodes of the binary tree are marked is sufficient. Let \( x_k \) be the maximum number of marked nodes in a tree, where \( k \) processes access the root. We claim that \( x_k \leq 2k - 1 \), which is true for \( k = 1 \) because a single process entering a splitter will always compute \( \text{stop} \). Now assume the inequality holds for \( 1, \ldots, k - 1 \). Not all \( k \) processes may exit the splitter with \text{left} (or \text{right}), i.e., \( k \leq k - 1 \) processes will turn left and \( k \leq \min(k - k, 1) \) turn right. The left and right children of the root are the roots of their subtrees, hence the induction hypothesis yields

\[
x_k = x_{k-1} + x_{k-2} + 1 \leq (2k - 1) + (2k - 1) + 1 \leq 2k - 1,
\]

concluding induction and proof. □
5.3. STORE & COLLECT

Similarly, the number of processes entering column $j$ is at most $k - j$. Hence, every process stops at the latest in row $k - 1$ and column $k - 1$ and the number of marked splitters is at most $k^2$. Thus, the step complexity of COLLECT is at most $O(k^2)$. Because the longest path in the splitter matrix is $2k$, the step complexity of STORE is $O(k)$.

Remarks:

- With a slightly more complicated argument, it is possible to show that the number of processes entering the splitter in row $i$ and column $j$ is at most $k - i - j$. Hence, it suffices to only allocate the upper left half (including the diagonal) of the $n \times n$ matrix of splitters.
- The binary tree algorithm can be made space efficient by using a randomized version of a splitter. Whenever returning left or right, a randomized splitter returns left or right with probability $1/2$. With high probability, it then suffices to allocate a binary tree of depth $O(\log n)$.
- Recently, it has been shown that with a considerably more complicated deterministic algorithm, it is possible to achieve $O(k)$ step complexity and $O(n^2)$ space complexity.

Chapter Notes

Already in 1965 Edsger Dijkstra gave a deadlock-free solution for mutual exclusion [Dij65]. Later, Maurice Herlihy suggested consensus-numbers [Her91], where he proved the “universality of consensus”, i.e., the power of a shared memory system is determined by the consensus-number. For this work, Maurice Herlihy was awarded the Dijkstra Prize in Distributed Computing in 2003. Peterson’s Algorithm is due to [PF77, Pet81], and adaptive collect was studied in the sequence of papers [MA95, AFG02, AL05, AKP06].

Bibliography


