Chapter 5

Shared Memory

5.1 Introduction

In distributed computing, various different models exist. So far, the focus of the course was 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 distributed computing such a system is called a Parallel Random Access Machine (PRAM).

A third major model is somewhere between these two extremes, the shared memory model. In a shared memory system, asynchronous processes (or processors) communicate via a common memory area of shared variables or registers:

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 R. Most systems do allow for advanced atomic read-modify-write (RMW) operations, for example:

  - test-and-set(R): t := R; R := 1; return t
  - fetch-and-add(R, x): t := R; R := R + x; return t
  - compare-and-swap(R, x, y): if R = x then R := y; return true; else return false; endif;
  - load-link(R)/store-conditional(R, x): Load-link returns the current value of the specified register R. A subsequent store-conditional to the same register will store a new value x (and return true) only if no updates have occurred to that register since the load-link. If any updates have occurred, the store-conditional is guaranteed to fail (and return false), even if the value read by the load-link has since been restored.

• The power of RMW operations can be measured with the so-called consensus-number. The consensus-number k of a RMW operation defines whether one can solve consensus for k processes. Test-and-set for instance has consensus-number 2 (one can solve consensus with 2 processes, but not 3), whereas the consensus-number of compare-and-swap is infinite. It can be shown that the power of a shared memory system is determined by the consensus-number (“universality of consensus.”) This insight has a remarkable theoretical and practical impact. In practice for instance, after this was known, hardware designers stopped developing shared memory systems supporting weak RMW operations.

• Many of the results derived in the message passing model have an equivalent in the shared memory model. Consensus for instance is traditionally studied in the shared memory model.

• Whereas programming a message passing system is rather tricky (in particular if fault-tolerance has to be integrated), programming a shared memory system is generally considered easier, as programmers are given access to global variables directly and do not need to worry about exchanging messages correctly. Because of this, even distributed systems which physically communicate by exchanging messages can often be programmed through a shared memory middleware, making the programmer’s life easier.

• We will most likely find the general spirit of shared memory systems in upcoming multi-core architectures. As for programming style, the multi-core community seems to favor an accelerated version of shared memory, transactional memory.

• From a message passing perspective, the shared memory model is like a bipartite graph: On one side you have the processes (the nodes) which pretty much behave like nodes in the message passing model (asynchronous, maybe failures). On the other side you have the shared registers, which just work perfectly (no failures, no delay).

5.2 Mutual Exclusion

A classic problem in shared memory systems is mutual exclusion. We are given a number of processes which occasionally need 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 the time is allowed to access the resource. More formally:

Definition 5.2 (Mutual Exclusion). We are given a number of processes, each executing the following code sections:

\(<\text{Entry}>\) \rightarrow \text{<Critical Section>} \rightarrow \text{<Exit>} \rightarrow \text{<Remaining Code>}

A mutual exclusion algorithm consists of code for entry and exit sections, such that 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 the same 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

```
<Initializations>
1. Repeat
2. \( r := \text{test-and-set}(R) \)
3. Until \( r = 0 \)
<Critical Section>
4. \[
5. R := 0
6. \]
<Exit>
5. \[ W_i := 0 \]
<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 \( W_i, \bar{W}_i, \Pi \), all initially 0.

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

\[
1. W_i := 1 \\
2. \Pi := 1 - i \\
3. \text{repeat until } \Pi = i \text{ or } W_{i - 1} = 0 \\
4. \text{[Critical Section]} \\
5. \Pi := 0 \\
6. \text{[Remainder Code]} \\
```

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 \( W_i \) by setting \( W_i := 1 \). Only if the other process is not interested (\( W_{i - 1} = 0 \)) 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 \( \Pi \). See Algorithm 24.

Algorithm 24 Mutual Exclusion: Peterson’s Algorithm

Initialization: Shared registers \( W_i, \bar{W}_i, \Pi \), all initially 0.

Code for process \( p_i, i = \{0, 1\} \)
```
1. \( W_i := 1 \)
2. \( \Pi := 1 - i \)
3. \text{repeat until } \Pi = i \text{ or } W_{i - 1} = 0 \\
4. \text{[Critical Section]} \\
5. \Pi := 0 \\
6. \text{[Remainder Code]} \\
```

Remarks:
- Note that line 3 in Algorithm 24 represents a “spinlock” or “busy-wait”, similarly to the lines 1-3 in Algorithm 23.

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

Proof. The shared variable \( \Pi \) elegantly grants priority to the process that passes \( \text{line 5} \). If both processes are competing, only process \( p_i \) gets the priority due to \( \Pi \), and can enter the critical section.
5.3 STORE & COLLECT

Remarks:
• Extending Peterson’s Algorithm to more than 2 processes can be done by a tournament tree, like in tennis. With \( n \) processes every process needs to win \( \log n \) matches before it can enter the critical section. More precisely, each process starts at the bottom level of a binary tree, and proceeds to the parent level if winning. Once winning the root of the tree it can enter the critical section. Thanks to the priority variables \( \Pi \) at each node of the binary tree, we inherit all the properties of Definition 5.2.

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 \( \text{op1} \) precedes an operation \( \text{op2} \) iff \( \text{op1} \) terminates before \( \text{op2} \) starts. An operation \( \text{op2} \) follows an operation \( \text{op1} \) iff \( \text{op1} \) precedes \( \text{op2} \).

Definition 5.5 (Collect). There are two operations: A \( \text{store}(v) \) by process \( p_i \) sets \( v \) to be 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 \( \text{sop} \) of \( p_i \) 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 \( \text{op} \) to be the number of accesses to registers in the shared memory. There is a trivial solution to the \( \text{collect} \) problem as shown by Algorithm 25.

Algorithm 25 Collect: Simple (Non-Adaptive) Solution

<table>
<thead>
<tr>
<th>Operation</th>
<th>\text{store}(val) (by process ( p_i )) :</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( R_i := \text{val} )</td>
</tr>
<tr>
<td>Operation ( \text{collect} ):</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>\text{for } i := 1 \text{ to } n \text{ do}</td>
</tr>
<tr>
<td>3</td>
<td>( V(p_i) := R_i ) \hspace{1cm} // read register ( R_i )</td>
</tr>
<tr>
<td>4</td>
<td>\text{end for}</td>
</tr>
</tbody>
</table>

Algorithm 26 Splitter Code

Shared Registers: \( X : \{\bot\} \cup \{1, \ldots, n\} \); \( Y : \text{boolean} \)

Initialization: \( X := \bot \); \( Y := \text{false} \)

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

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

Definition 5.6 (Splitter). A \( \text{split} \) is a synchronization primitive with the following characteristics. A process entering a \( \text{split} \) exits with either \( \text{stop} \), \( \text{left} \), or \( \text{right} \). If \( k \) processes enter a \( \text{split} \), at most one \( \text{process} \) 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 \( \text{split} \), then it obtains \( \text{stop} \), and if two or more processes enter the \( \text{split} \), then there is at most one process obtaining \( \text{stop} \) and there are two processes that obtain
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{true}. Next, assume that \( i \) is the last process that sets \( X = i \). If \( i \) does not return \text{true}, it will find \( X = i \) in line 6 and therefore return \text{stop}. Hence, there is always a process that does not return \text{true}. It remains to show that at most \( 1 \) process returns \text{stop}. For the sake of contradiction, assume \( p_1 \) and \( p_2 \) are two processes that return \text{stop} and assume that \( p_1 \) sets \( X := i \) before \( p_2 \) sets \( X := j \). Both processes need to check whether \( Y \) is \text{true} before one of them sets \( Y := \text{false} \). 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 \( p_1 \) arrives at line 6, \( X \neq i \). \( p_1 \) and maybe some other processes have overwritten \( X \) by then). Therefore, \( p_2 \) does not return \text{stop} and we get a contradiction to the assumption that both \( p_1 \) and \( p_2 \) return \text{stop}.

5.3.3 Binary Splitter Tree

Assume that we are given \( 2^k - 1 \) splitters and that for every splitter \( S \), there is an additional shared variable \( Z_S : \mathbb{S} \cup [1, \ldots, n] \) that is initialized to \( \perp \) and an additional shared variable \( M_S : \text{boolean} \) that is initialized to \text{false}. We call a splitter \( S \) marked if \( M_S = \text{false} \). The \( 2^k - 1 \) splitters are arranged in a complete binary tree of height \( k - 1 \). 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 \text{stopping at} 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.
Remarks:

- The step complexities of Algorithm 27 are very good. Clearly, the step complexity of the collect operation is asymptotically optimal. In order for the algorithm to work, we however need to allocate the memory for the complete binary tree of depth $n - 1$. The space complexity of Algorithm 27 therefore is exponential in $n$. We will next see how to obtain a polynomial space complexity at the cost of a worse collect step complexity.

### 5.3.4 Splitter Matrix

Instead of arranging splitters in a binary tree, we arrange $n^2$ splitters in an $n \times n$ matrix as shown in Figure 5.2. The algorithm is analogous to Algorithm 27. The matrix is entered at the top left. If a process receives left, it next visits the splitter in the next row of the same column. If a process receives right, it next visits the splitter in the next column of the same row. Clearly, the space complexity of this algorithm is $O(n^2)$. The following theorem gives bounds on the step complexities of store and collect.

**Theorem 5.9.** Let $k$ be the number of participating processes. The step complexity of the first store of a process $p_i$ is $O(k)$, and the step complexity of every additional store of $p_i$ is $O(1)$, and the step complexity of collect is $O(k^2)$.

**Proof.** Let the top row be row 0 and the left-most column be column 0. Let $x_i$ be the number of processes entering a splitter in row $i$. By induction on $i$, we show that $x_i \leq k - i$. Clearly, $x_0 \leq k$. Let us therefore consider the case $i > 0$. Let $j$ be the largest column such that at least one process visits the splitter in row $i - 1$ and column $j$. By the properties of splitters, not all processes entering the splitter in row $i - 1$ and column $j$ obtain left. Therefore, not all processes entering a splitter in row $i - 1$ move on to row $i$. Because at least one process stays in every row, we get that $x_i \leq k - i$. 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 [Dj65]. 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 [P77, Fe84], and adaptive collect was studied in the sequence of papers [MA95, AFG02, AL05, AKP+06].

### Bibliography
