Relaxed Memory-Consistency Models

[§9.1] In small multiprocessors, sequential consistency can be implemented relatively easily. However, this is not true for large multiprocessors. Why?

This is not the only problem with sequential consistency. Let’s see how it prevents performance optimizations that are common in compilers.

Compilers allocate variables to registers for better performance, but this is not possible if sequential consistency is to be preserved.

In the example below, assume that A and B are initialized to 0.

<table>
<thead>
<tr>
<th>Without register allocation</th>
<th>After register allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1 )</td>
<td>( P_1 )</td>
</tr>
<tr>
<td>A = 1</td>
<td>r1 = 0</td>
</tr>
<tr>
<td>u = B</td>
<td>A = 1</td>
</tr>
<tr>
<td>B = 1</td>
<td>B = 1</td>
</tr>
<tr>
<td>v = A</td>
<td>u = r1</td>
</tr>
<tr>
<td></td>
<td>v = r2</td>
</tr>
<tr>
<td></td>
<td>B = r1</td>
</tr>
<tr>
<td></td>
<td>A = r2</td>
</tr>
</tbody>
</table>

Note that this use of registers is valid for a sequential program.

After executing the code without register allocation, what final values are possible for \( u \) and \( v \)?

After executing the code with register allocation, what final values are possible for \( u \) and \( v \)?

So, mere usage of processor registers causes a result not possible with sequential consistency.
How can we attack this problem? Well, ...

Basically, what we want to do is develop models where one operation doesn’t have to complete before another one is issued, but the system assures that operations don’t become visible out of program order.

Clearly, we must circumvent sequential consistency if we want our programs to perform acceptably.

- We can preserve sequential consistency, but *overlap operations* as much as possible.
  E.g., we can prefetch data, or use more multithreading.
  But sequential consistency is preserved, so the compiler cannot reorder operations.

- We can *allow the compiler to reorder operations* as long as we are sure that sequential consistency will not be violated in the results.
  Memory operations can be executed out of order, as long as they become visible to other processors in program order.
  For example, we can execute instructions speculatively, and simply not commit the results if those instructions should not have been executed.

- We can *relax the consistency model*, abandoning the guarantee of sequential consistency, but still retain semantics that are intuitive.

The reason that relaxed consistency models are useful is that most of the sequential constraints in a program are not really necessary for it to give intuitively correct results.

\[
\begin{align*}
P_1 & \\
1a. & \quad A = 1 \\
1b. & \quad B = 1 \\
& \quad P_2 \\
2a. & \quad \text{while (flag == 0)}; \\
2b. & \quad u = A
\end{align*}
\]
1c. \texttt{flag = 1} \hspace{1cm} 2c. \texttt{v = B}

If the code were to execute sequentially, statement 1a would have to precede 1b, which would have to precede 1c; ditto for 2a–2c.

What precedence relationships among the statements are really necessary?

**A continuum of consistency models**

Sequential consistency is one data point on what should be required of a programming model.

Let us introduce a way of diagramming consistency models. Suppose that—

- The value of a particular memory word in processor 2’s local memory is 0.
- Then processor 1 writes the value 1 to that word of memory. Note that this is a remote write.
- Processor 2 then reads the word. But, being local, the read occurs quickly, and the value 0 is returned.

What’s wrong with this? Processor 2 has retrieved an out-of-date value.

This situation can be diagrammed like this (the horizontal axis represents time):

$$
\begin{align*}
P_1: & \quad W(x)1 \\
P_2: & \quad R(x)0
\end{align*}
$$

Depending upon how the program is written, it may or may not be able to tolerate a situation like this.

But, in any case, the programmer must understand what can happen when memory is accessed in a DSM system.
One consistency model is even stricter than sequential consistency.

**Strict consistency**

**Strict consistency:** Any read to a memory location \( x \) returns the value stored by the most recent write operation to \( x \).

This definition implicitly assumes the existence of a global clock, so that the determination of “most recent” is unambiguous.

Strict consistency is supported by uniprocessors, but it is generally impossible to implement in DSM systems.

In a DSM system, a read from a nonlocal memory location, or a write to a nonlocal memory location, requires sending a message.

This message requires a finite time to travel to and from the node where the memory is located. During that time, some other processor might change the value in the memory location.

Regardless of how efficient a message system is, it cannot cause messages to travel faster than the speed of light.

If the difference in the two access times is, say, one ns., and the nodes are 3 m. apart, the signal would have to travel at 10 times the speed of light to return the most recent value.

Fortunately, strict consistency is rarely necessary.

To summarize, with strict consistency,

- all writes are instantaneously visible to all processes;
- all subsequent reads see the new value of the write, no matter how soon after the write the read is done, and no matter where the process executing the read is located.

**Sequential consistency**

Strict consistency isn’t really necessary to write parallel programs.
Parallel programs shouldn’t make any assumptions about the relative speeds of the processes, or how their actions would interleave in time.

Counting on two events within one process to happen so quickly that another process won’t have time to do something in between is asking for trouble.

Let us weaken the requirements we impose on the memory so that the resulting model is realizable.

**Sequential consistency:** The result of any execution is the same as if

- the memory operations of all processors were executed in some sequential order, and
- the operations of each individual processor appear in this sequence in the order specified by its program.

What’s the difference between strict and sequential consistency?

- In sequential consistency, the temporal ordering of events does not matter.

- All that is required is that the processors see the same ordering of memory operations (regardless of whether this is the order that the operations actually occurred).

So, with sequential consistency we don’t have to worry about providing “up to the nanosecond” results to all of the processors.

The example below shows the difference between strict and sequential consistency. The two sequences of operations are equally valid.

Note that a read from \( P_2 \) is allowed to return an out-of-date value (because it has not yet “seen” the previous write).

\[
\begin{align*}
P_1 &: W(x)1 \\
P_2 &: R(x)0 \quad R(x)1 \\
P_1 &: W(x)1 \\
P_2 &: R(x)1 \quad R(x)1
\end{align*}
\]
From this we can see that running the same program twice in a row in a system with sequential consistency may not give the same results.

**Causal consistency**

The next step in weakening the consistency constraints is to distinguish between events that are potentially *causally* connected and those that are not.

Two events are causally related if one can influence the other.

\[
P_1: \ W(x)1 \\
P_2: \ R(x)1 \ W(y)2
\]

Here, the write to \( x \) could influence the write to \( y \), because

On the other hand, without the intervening read, the two writes would not have been causally connected:

\[
P_1: \ W(x)1 \\
P_2: \ W(y)2
\]

The following pairs of operations are potentially causally related:

- A read followed by a later write.
- A write followed by a later read to the same location.
- The transitive closure of the above two types of pairs of operations.

Operations that are not causally related are said to be *concurrent*.

*Exercise:* Come up with a set of three processes and four writes, one of which is causally related to *all* the other writes, two of which are causally connected to *some* of the other writes, and one of which is not causally related to *any* of the other writes.
**Causal consistency:** *Writes that are potentially causally related must be seen in the same order by all processors.*

*Concurrent writes may be seen in a different order by different processors.*

Here is a sequence of events that is allowed with a causally consistent memory, but disallowed by a sequentially consistent memory:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$:</td>
<td>$W(x)1$</td>
<td>$W(x)3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_2$:</td>
<td>$R(x)1$</td>
<td>$W(x)2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_3$:</td>
<td>$R(x)1$</td>
<td>$R(x)3$</td>
<td>$R(x)2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_4$:</td>
<td>$R(x)1$</td>
<td>$R(x)2$</td>
<td>$R(x)3$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why is this not allowed by sequential consistency?

Why is this allowed by causal consistency?

What is the violation of causal consistency in the sequence below?

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$:</td>
<td>$W(x)1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_2$:</td>
<td>$R(x)1$</td>
<td>$W(x)2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_3$:</td>
<td>$R(x)2$</td>
<td>$R(x)1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_4$:</td>
<td>$R(x)1$</td>
<td>$R(x)2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Without the $R(x)1$ by $P_2$, this sequence would’ve been legal. Implementing causal consistency requires the construction of a dependency graph, showing which operations depend on which other operations.
PRAM consistency

Causal consistency requires that all processes see causally related writes from all processors in the same order.

The next step is to relax this requirement, to require only that writes from the same processor be seen in order. This gives pipelined-RAM (PRAM) consistency.

PRAM consistency: Writes performed by a single process are received by all other processors in the order in which they were issued.

Writes from different processors may be seen in a different order by different processors.

PRAM consistency is so named because writes can be pipelined; that is, a processor does not have to stall waiting for a write to be completed before starting the next one.

PRAM consistency would permit this sequence that we saw violated causal consistency:

\[
\begin{align*}
P_1: & \quad W(x)1 \\
P_2: & \quad R(x)1 \quad W(x)2 \\
P_3: & \quad R(x)2 \quad R(x)1 \\
P_4: & \quad R(x)1 \quad R(x)2
\end{align*}
\]

Another way of looking at this model is that all writes generated by different processors are considered to be concurrent.

Sometimes PRAM consistency can lead to counterintuitive results.

\[
\begin{align*}
P_1: & \quad P_2:
\end{align*}
\]
\begin{verbatim}
a := 0;  
:  
a := 1;  
if b = 0 then kill(p2);  
\end{verbatim}

\begin{verbatim}
b := 0;  
:  
b := 1;  
if a = 0 then kill(p1);  
\end{verbatim}

At first glance, it seems that no more than one process should be killed. With PRAM consistency, however, it is possible for both to be killed.

\textit{Processor consistency}

Processor consistency is very similar to PRAM consistency, but it has one additional condition: memory coherence.

Memory \textit{coherence} requires that writes to the \textit{same location} be viewed in the same order by all the processors.

Writes to different locations need not be seen in the same order by different processors.

\textbf{Processor consistency} = \textit{PRAM} consistency + memory coherence.

\textit{Weak ordering}

PRAM and processor consistency are still stronger than necessary for many programs, because they require that writes originating in a single processor be seen in order everywhere.

But it is not always necessary for other processors to see writes in order—or even to see all writes, for that matter.

Suppose a processor is in a tight loop in a critical section, reading and writing variables.

Other processes aren’t supposed to touch these variables until the process exits its critical section.
Under PRAM consistency, the memory has no way of knowing that other processes don’t care about these writes, so it has to propagate all writes to all other processors in the normal way.

To relax our consistency model further, we have to divide memory operations into two classes and treat them differently.

- Accesses to synchronization variables are sequentially consistent.
- Accesses to other memory locations can be treated as concurrent.

This strategy is known as weak ordering.

With weak ordering, we don’t need to propagate accesses that occur during a critical section.

We can just wait until the process exits its critical section, and then—

- make sure that the results are propagated throughout the system, and
- stop other actions from taking place until this has happened.

Similarly, when we want to enter a critical section, we need to make sure that all previous writes have finished.

These constraints yield the following definition:

**Weak ordering:** A memory system exhibits weak ordering iff—

1. Accesses to synchronization variables are sequentially consistent.
2. No access to a synchronization variable can be performed until all previous writes have completed everywhere.
3. No data access (read or write) can be performed until all previous accesses to synchronization variables have been performed.
Thus, by doing a synchronization before reading shared data, a process can be assured of getting the most recent values.

Note that this model does not allow more than one critical section to execute at a time, even if the critical sections involve disjoint sets of variables.

This model puts a greater burden on the programmer, who must decide which variables are synchronization variables.

Weak ordering says that memory does not have to be kept up to date between synchronization operations.

This is similar to how a compiler can put variables in registers for efficiency’s sake. Memory is only up to date when these variables are written back.

If there were any possibility that another process would want to read these variables, they couldn’t be kept in registers.

This shows that processes can live with out-of-date values, provided that they know when to access them and when not to.

The following is a legal sequence under weak ordering. Can you explain why?

\[
\begin{array}{lll}
P_1: & W(x)1 & W(x)2 & S \\

P_2: & & R(x)2 & R(x)1 & S \\

P_3: & & R(x)1 & R(x)2 & S \\
\end{array}
\]

Here’s a sequence that’s illegal under weak ordering. Why?

\[
\begin{array}{lll}
P_1: & W(x)1 & W(x)2 & S \\

P_2: & & S & R(x)1 \\
\end{array}
\]
Release consistency

Weak ordering does not distinguish between entry to critical section and exit from it.

Thus, on both occasions, it has to take the actions appropriate to both:

- making sure that all locally initiated writes have been propagated to all other memories, and
- making sure that the local processor has seen all previous writes anywhere in the system.

If the memory could tell the difference between entry and exit of a critical section, it would only need to satisfy one of these conditions.

Release consistency provides two operations:

- *acquire* operations tell the memory system that a critical section is about to be entered.
- *release* operations say a c. s. has just been exited.

It is possible to acquire or release a single synchronization variable, so more than one critical section can be in progress at a time.

When an acquire occurs, the memory will make sure that all the local copies of shared variables are brought up to date.

When a release is done, the shared variables that have been changed are propagated out to the other processors.

But—

- doing an acquire does not guarantee that locally made changes will be propagated out immediately.
- doing a release does not necessarily import changes from other processors.
Here is an example of a valid event sequence for release consistency (A stands for “acquire,” and Q for “release” or “quit”):

\[
\begin{align*}
P_1: & \quad A(L) \quad W(x)1 \quad W(x)2 \quad Q(L) \\
P_2: & \quad A(L)R(x)2 \quad Q(L) \\
P_3: & \quad R(x)1
\end{align*}
\]

Note that since \( P_3 \) has not done a synchronize, it does not necessarily get the new value of \( x \).

**Release consistency:** A system is release consistent if it obeys these rules:

1. Before an ordinary access to a shared variable is performed, all previous acquires done by the process must have completed.
2. Before a release is allowed to be performed, all previous reads and writes done by the process must have completed.
3. The acquire and release accesses must be processor consistent.

If these conditions are met, and processes use acquire and release properly, the results of an execution will be the same as on a sequentially consistent memory.

**Summary:** Strict consistency is impossible. Sequential consistency is possible, but costly. The model can be relaxed in various ways. Consistency models not using synchronization operations:
<table>
<thead>
<tr>
<th>Type of consistency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strict</td>
<td>All processes see absolute time ordering of all shared accesses.</td>
</tr>
<tr>
<td>Sequential</td>
<td>All processes see all shared accesses in same order.</td>
</tr>
<tr>
<td>Causal</td>
<td>All processes see all causally related shared accesses in the same order.</td>
</tr>
<tr>
<td>Processor</td>
<td><strong>PRAM</strong> consistency + memory coherence</td>
</tr>
<tr>
<td><strong>PRAM</strong></td>
<td>All processes see writes from each processor in the order they were initiated. Writes from different processors may not be seen in the same order.</td>
</tr>
</tbody>
</table>

Consistency models using synchronization operations:

<table>
<thead>
<tr>
<th>Type of consistency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>Shared data can only be counted on to be consistent after a synchronization is done.</td>
</tr>
<tr>
<td>Release</td>
<td>Shared data are made consistent when a critical region is exited.</td>
</tr>
</tbody>
</table>

The following diagram contrasts various forms of consistency.

![Diagram](https://example.com/diagram.png)