Concurrent Systems
Exercise 03 – Memory, Atomicity, Consistency

Stefan Reif

November 28, 2016
Agenda

Memory Models

Sequential Consistency

C11/C++11 Atomic Types

C/C++ Memory Consistency Models

Thread Fences

Assignment 3
What is a memory model?

- **Formal definition of memory behavior**
  - More or less trivial for sequential programs
  - Complex for parallel programs

- **Crucial for application correctness**
  - Particularly for parallel programs

- **Memory models are created by ...**
  - Language designers
  - Hardware developers
  - Service providers
Sequential Consistency

Def. Sequential Consistency \hspace{1cm} \text{Lamport, 1979}

“[...] the result of any execution is the same as if the operations of all the 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.”

Benefits of SC
- Result is equivalent to a sequential execution
  \rightarrow Pseudo-parallelism?
- Global order of actions
- Analyzability

Problems of SC
- Even simple compiler optimizations violate SC
Why don’t we always want SC?

- Modern Platforms are not WYSIWYG\(^1\)
  - Optimizing Compiler, Assembler, Linker, Hardware
  - Every link in the tool-chain can reorder operations
    - We nearly always want that ...
    - ... but it can break parallel code
    - It does not matter which link breaks our code

- SC makes optimizations hard

More Reordering

SC \[\rightarrow\] Analyzability \[\rightarrow\] Bugs?
SC \[\rightarrow\] Correctness \[\rightarrow\] Performance?

\(^1\)What You See Is What You Get
Sequential Consistency – Data Race Free

SC-DRF
- Distinguish synchronizing actions from non-synchronizing actions
  - Guarantee SC for programs without data races
  - Undefined behavior in case of data race
  - Most code parts allow optimization
- Synchronizing actions provide SC
  - e.g. mutex_lock, atomic_fetch_add, ...
- Non-synchronizing actions can cause data races

Def. *Data Race*  

“Two expression evaluations *conflict* if one of them modifies a memory location and the other one reads or modifies the same memory location.”

“The execution of a program contains a *data race* if it contains two conflicting actions in different threads, at least one of which is not atomic, and neither happens before the other.”
Sequential Consistency – Data Race Free

thread1

thread2

(sr cs-ex3 (November 28, 2016)) Sequential Consistency
Modern C/C++ offers atomic types

- C ⇒ _Atomic qualifier
- C++ ⇒ std::atomic<type> template
- Intentionally compatible:

```c
#ifdef __cplusplus
#define _Atomic(type) std::atomic<type>
#endif
```

Well-defined semantics

- Compiler guarantees atomicity
- *unlike volatile* ...

Portable

- Actual run-time properties still depend on the hardware
- Lock-freedom is not guaranteed
C11 Atomic Operations

Atomic operations

```c
#include <stdatomic.h>

atomic_store(AT*, T);
atomic_load(AT*);
atomic_exchange(AT*, T);
atomic_compare_exchange_strong(AT*, T*, T);
atomic_compare_exchange_weak(AT*, T*, T);
atomic_fetch_add(AT*, T);
atomic_fetch_sub(AT*, T);
atomic_fetch_or(AT*, T);
atomic_fetch_xor(AT*, T);
atomic_fetch_and(AT*, T);
```
C11 Atomic Operations

Atomic operations with explicit memory order parameter

```c
#include <stdatomic.h>
atomic_store_explicit(AT*, T, MO);
atomic_load_explicit(AT*, MO);
atomic_exchange_explicit(AT*, T, MO);
atomic_compare_exchange_strong_explicit(AT*, T*, T, MO, MO);
atomic_compare_exchange_weak_explicit(AT*, T*, T, MO, MO);
atomic_fetch_add_explicit(AT*, T, MO);
atomic_fetch_sub_explicit(AT*, T, MO);
atomic_fetch_or_explicit(AT*, T, MO);
atomic_fetch_xor_explicit(AT*, T, MO);
atomic_fetch_and_explicit(AT*, T, MO);
```
C/C++ Memory Order Parameters

1. sequential-consistent
   - memory_order_seq_cst

2. acquire-release
   - memory_order_acquire
   - memory_order_release
   - memory_order_acq_rel

3. consume-release
   - memory_order_consume
   - memory_order_release

4. relaxed
   - memory_order_relaxed
C/C++ Memory Order Parameters

1. sequential-consistent
   - memory_order_seq_cst

2. acquire-release
   - memory_order_acquire
   - memory_order_release
   - memory_order_acq_rel

3. consume-release
   - memory_order_consume
   - memory_order_release

4. relaxed
   - memory_order_relaxed
memory_order_seq_cst

- Strongest consistency model
- Implicit consistency model
  - All operations without `_explicit`
- Semantics: SC-DRF
  - All operations with `memory_order_seq_cst` are sequentially consistent
  - Non-atomic variables can cause data races
memory_order_{acquire, release, acq_rel}

- Happens-before relation
  - Intra-thread: trivial
  - Inter-thread: If $S_A$ release-stores a value $v$ in $A$ and $L_A$ acquire-loads that value, then $S_A$ happens before $L_A$.
  - $S_A$ also happens before $L_A$ if $L_A$ reads a later value $v'$ of $A$
    $\rightarrow$ modification order
  - Transitive relation

- Acquire/Release consistency
  - Each operation sees side effects of all operations that happened before

- Partial Ordering
  - No global sequence of operations
  - Synchronization per variable
  - Concurrent operations are possible
**memory_order_{acquire,release,acq_rel}**  \( (2) \)

- **acquire → load**
  - *fetch-all-data*
  - Side effects after `load-acquire` remain afterwards
  - Implicit in `thrd_join`, `mtx_lock`

- **release → store**
  - *push-all-data*
  - Side effects before `store-release` remain before
  - Implicit in `thrd_exit`, `mtx_unlock`

- **acq_rel → load/store**
  - e.g. `atomic_fetch_and_*`, `atomic_compare_exchange_*`
memory_order_{acquire,release,acq_rel} (3)
memory_order_{acquire, release, acq_rel}
memory_order_\{acquire,release,acq_rel\} (3)
memory_order_{acquire,release,acq_rel} (3)
memory_order_{acquire, release, acq_rel} (4)
memory_order_{acquire,release,acq_rel} (4)

thread1

LOAD

✓CAS

thread2

LOAD

acquire
release

acquire
release

acquire
release

CAS✗

CAS✓
Example:

```c
int data; __Atomic(bool) avail;
thread1() {
    data = createdata(); // not atomic
    atomic_store_explicit(&avail, 1, memory_order_release);
}
thread2() {
    if (atomic_load_explicit(&avail, memory_order_acquire)) {
        int mycopy = data; use(mycopy);
    }
}
```
Example:

```c
_Atomic(foo *) data;
thread1() {
    foo *d = createdata(); // not atomic
    atomic_store_explicit(&data, d,
                          memory_order_release);
}
thread2() {
    foo *d = atomic_load_explicit(&data,
                                   memory_order_acquire);
    if (d)
        use(d);
}```
memory_order_\{seq_cst vs. acquire/release\}

- Difference between seq_cst and acquire/release

```c
X = 1;
Y = 1;
while (!X);
A = Y;
while (!Y);
B = X;
```

- Possible Outcome

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>seq_cst</th>
<th>acquire/release</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Difference between seq_cst and acquire/release

\[
\begin{align*}
X &= 1; \\
\text{while} &\text{ (!X);} \\
A &= Y; \\
Y &= 1; \\
\text{while} &\text{ (!Y);} \\
B &= X;
\end{align*}
\]

Possible Outcome

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>seq_cst</th>
<th>acquire/release</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
memory_order_{seq_cst vs. acquire/release}

- Difference between seq_cst and acquire/release

X = 1;
Y = 1;

while (!X);
A = Y;

while (!Y);
B = X;

- Possible Outcome

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>seq_cst</th>
<th>acquire/release</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Difference between seq_cst and acquire/release

X = 1;
while (!X);
A = Y;

Y = 1;
while (!Y);
B = X;

Possible Outcome

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>seq_cst</th>
<th>acquire/release</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>
memory_order_relaxed

- Weakest memory order parameter
  - Meaningful when additional mechanisms synchronize

- Ensures atomicity
  - No synchronization
  - Lock-freedom not guaranteed

- No reordering constraints
  - Except for thread fences ...
  - Dangerous to use

- Performance improvement?
```c
_Atomic(unsigned int) counter;

search() {
  while (...) {
    if (...) 
      atomic_fetch_add_explicit(&counter, 1, 
        memory_order_relaxed);
  }
  thrd_exit(...);
}

main() {
  for (...) thrd_create(..., search, ...);
  for (...) thrd_join(...);
  unsigned int total = atomic_load_explicit(&counter, 
    memory_order_relaxed);
}
```
```c
_Atomic(unsigned int) counter;

search() {
    while (...) {
        if (...) atomic_fetch_add_explicit(&counter, 1, memory_order_relaxed);
    }
    thrd_exit(...);
}

main() {
    for (...) thrd_create(..., search, ...);
    for (...) thrd_join(...);
    unsigned int total = atomic_load_explicit(&counter, memory_order_relaxed);
}
```
Enforces additional memory ordering guarantees
- Often better: use stronger memory order parameter at critical operation

Parameter: memory order
- memory_order_acquire
- memory_order_release
- memory_order_acq_rel
- memory_order_seq_cst

Improves consistency of relaxed-atomic operations
- Useful for library functions
atomic_thread_fence

thread1 -> relaxed

A

relaxed

thread2
atomic_thread_fence

thread1

thread2

relaxed

A

X

relaxed

X
atomic_thread_fence
atomic_thread_fence

thread1
release
relaxed

thread2
relaxed
acquire
atomic_thread_fence
int data; atomic_bool avail;
thread1() {
    data = createdata(); // not atomic
atomic_store_explicit(&avail, 1,
    memory_order_relaxed);
}
thread2() {
    if (atomic_load_explicit(&avail,
        memory_order_relaxed)) {
        int mydata = data; use(mydata);
    }
}
int data; atomic_bool avail;
thread1() {
    data = createdata(); // not atomic
    atomic_thread_fence(memory_order_release);
    atomic_store_explicit(&avail, 1,
        memory_order_relaxed);
}
thread2() {
    if (atomic_load_explicit(&avail,
        memory_order_relaxed)) {
        atomic_thread_fence(memory_order_acquire);
        int mydata = data; use(mydata);
    }
}
Assignment 3

- Implement lock algorithms
  - Lock algorithms are discussed in the lecture

- Test all lock algorithms

- Evaluate all lock algorithms
  - Vary the degree of contention
  - Check multiple back-off algorithms

- Use a model checker for concurrent data structures
  - CDSChecker → plrg.eecs.uci.edu/software_page/42-2/
  - Write a unit test for at least one lock algorithm
Implement a simple actor library
- Actors send requests asynchronously
- Each actor owns a worker thread
- Requests are sequentialized implicitly

Shortcomings
- No actor states and mutation
- No actor migration
- ...

Test your actor library
- Write test cases for your actor
- Can you re-use the lock test cases?

Evaluate your actor library
- Compare your actor implementation against lock-based synchronization
- Under which conditions are actors a better choice?