Lecture 9 — C++ Atomics, Compiler Hints, Restrict

Patrick Lam & Jeff Zarnett
p.lam@ece.uwaterloo.ca, jzarnett@uwaterloo.ca

Department of Electrical and Computer Engineering
University of Waterloo

December 3, 2017

Locks and Atomics

Atomics are a lower-overhead alternative to locks as long as you're doing suitable operations.

Remember that what we wanted sometimes with locks and mutexes and all that is that operations are indivisible.

Ex: an update to a variable doesn't get interfered with by another update.

Remember the key idea is: an atomic operation is indivisible.
Other threads see state before or after the operation; nothing in between.

About C++ atomics

You can use the default `std::memory_order`.
(= sequential consistency)

Don't use relaxed atomics unless you're an expert!

Different Memory Options

- memory_order_acquire
- memory_order_release
- memory_order_acq_rel
- memory_order_consume
- memory_order_relaxed
- memory_order_seq_cst

Really, don’t use C++ relaxed atomics!

C++ atomics: Key Idea

An *atomic operation* is indivisible.

Other threads see state before or after the operation, nothing in between.

Simplest: `atomic_flag`

```
#include <atomic>

atomic_flag f = ATOMIC_FLAG_INIT;
```

Represents a boolean flag.

Operations on `atomic_flag`

Can clear, and can test-and-set:

```
#include <atomic>

atomic_flag f = ATOMIC_FLAG_INIT;
int foo() {
  f.clear();
  if (f.test_and_set()) {
    // was true
  }
}
```

`test_and_set`: atomically sets to true,
returns previous value.

No assignment (=) operator.

Although I guess in C++ you could define one if you wanted.

This is kind of a dangerous thing about C++.

If in C you see a line of code like `z = x + y;` you can have a pretty good idea about what it does and you can infer that there's some sort of natural meaning to the + operator there, like addition or concatenation.

In C++, however, this same line of code tells you nothing unless you know...
(1) the type of `x`,
(2) the type of `y`, and
(3) how the + operator is defined on those two operands *in that order*.

But I'm digressing.

Using more general C++ atomics

Declaring them:

```
#include <atomic>

atomic<int> x;
```

Libary's implementation:

- on small types, lock-free operations;
- on large types, mutexes.

What to do with Atomics

Kinds of operations:

- reads
- writes
- read-modify-write (RMW)

Reads and Writes

C++ has syntax to make these all transparent:

```
#include <atomic>
#include <iostream>

std::atomic<int> ai;
int i;

int main() {
    ai = 4;
    i = ai;
    ai = i;
    std::cout << i;
}
```

Can also use `i = ai.load()` and `ai.store(i)`.

Read-Modify-Write (RMW)

Consider `ai++`.

This is really

```
tmp = ai.read(); tmp++; ai.write(tmp);
```

Hardware can do that atomically.

Other RMWs: `+-`, `&=`, `etc`, `compare-and-swap`

more info:
`http://preshing.com/20130618/atomic-vs-non-atomic-operations/`

Three Address Code

- An intermediate code used by compilers for analysis and optimization.
- Statements represent one fundamental operation—we can consider each operation atomic.
- Statements have the form:
  $$result := operand_1\ operator\ operand_2$$
- Useful for reasoning about data races,
  and easier to read than assembly.
  (separates out memory reads/writes).

GIMPLE

- GIMPLE is the three address code used by `gcc`.
- To see the GIMPLE representation of your code use the `-fdump-tree-gimple` flag.
- To see all of the three address code generated by the compiler use `-fdump-tree-all`. You'll probably just be interested in the optimized version.
- Use GIMPLE to reason about your code at a low level without having to read assembly.

Branch Prediction Hints

As seen earlier in class, `gcc` allows you to give branch prediction hints by calling this builtin function:

```
long __builtin_expect (long exp, long c)
```

The expected result is that `exp` equals `c`.

Compiler reorders code & tells CPU the prediction.

The `restrict` Keyword

A new feature of C99: “The restrict type qualifier allows programs to be written so that translators can produce significantly faster executables.”

- To request C99 in `gcc`, use the `-std=c99` flag.

`restrict` means: you are promising the compiler that the pointer will never alias (another pointer will not point to the same data) for the lifetime of the pointer.

Take the Vow

I, [insert your name], a PROFESSIONAL or AMATEUR [circle one] programmer recognize that there are limits to what a compiler can do. I certify that, to the best of my knowledge, there are no magic elves or monkeys in the compiler which through the forces of fairy dust can always make code faster. I understand that there are some problems for which there is not enough information to solve. I hereby declare that given the opportunity to provide the compiler with sufficient information, perhaps through some key word, I will gladly use said keyword and not bitch and moan about how "the compiler should be doing this for me."

In this case, I promise that the pointer declared along with the restrict qualifier is not aliased. I certify that writes through this pointer will not effect the values read through any other pointer available in the same context which is also declared as restricted.

* Your agreement to this contract is implied by use of the restrict keyword ;)

Example of `restrict` (1)

Pointers declared with `restrict` must never point to the same data.

From Wikipedia:

```
void updatePtrs(int* ptrA, int* ptrB, int* val) {
    *ptrA += *val;
    *ptrB += *val;
}
```

Would declaring all these pointers as `restrict` generate better code?

Example of `restrict` (2)

Let's look at the GIMPLE:

```
void updatePtrs(int* ptrA, int* ptrB, int* val) {
 D.1609 = *ptrA;
 D.1610 = *val;
 D.1611 = D.1609 + D.1610;
 *ptrA = D.1611;
 D.1612 = *ptrB;
 D.1610 = *val;
 D.1613 = D.1612 + D.1610;
 *ptrB = D.1613;
}
```

- Could any operation be left out if all the pointers didn't overlap?

Example of `restrict` (3)

```
void updatePtrs(int* ptrA, int* ptrB, int* val) {
 D.1609 = *ptrA;
 D.1610 = *val;
 D.1611 = D.1609 + D.1610;
 *ptrA = D.1611;
 D.1612 = *ptrB;
 D.1610 = *val;
 D.1613 = D.1612 + D.1610;
 *ptrB = D.1613;
}
```

- If `ptrA` and `val` are not equal, you don't have to reload the data on **line 7**.
- Otherwise, you would: there might be a call

  `updatePtrs(&x, &y, &x);`

Example of `restrict` (4)

Hence, this markup allows optimization:

```
void updatePtrs(int* restrict ptrA,
                int* restrict ptrB,
                int* restrict val)
```

Note: you can get the optimization by just declaring `ptrA` and `val` as `restrict`; `ptrB` isn't needed for this optimization

Summary of `restrict`

- Use `restrict` whenever you know the pointer will not alias another pointer (also declared `restrict`)

It's hard for the compiler to infer pointer aliasing information; it's easier for you to specify it.

$\Rightarrow$ compiler can better optimize your code (more perf!)

Caveat: don't lie to the compiler, or you will get undefined behaviour.

Aside: `restrict` is not the same as `const`. `const` data can still be changed through an alias.