M.3 — Move constructors and move assignment

ALEX AUGUST 26, 2021

In lesson M.1 -- Intro to smart pointers and move semantics, we took a look at std::auto_ptr, discussed the desire for move semantics, and took a look at some of the downsides that occur when functions designed for copy semantics (copy constructors and copy assignment operators) are redefined to implement move semantics.

In this lesson, we'll take a deeper look at how C++11 resolves these problems via move constructors and move assignment.

Copy constructors and copy assignment

First, let's take a moment to recap copy semantics.

Copy constructors are used to initialize a class by making a copy of an object of the same class. Copy assignment is used to copy one class object to another existing class object. By default, C++ will provide a copy constructor and copy assignment operator if one is not explicitly provided. These compiler-provided functions do shallow copies, which may cause problems for classes that allocate dynamic memory. So classes that deal with dynamic memory should override these functions to do deep copies.

Returning back to our Auto_ptr smart pointer class example from the first lesson in this chapter, let's look at a version that implements a copy constructor and copy assignment operator that do deep copies, and a sample program that exercises them:

```cpp
template<class T>
class Auto_ptr3
{
 T* m_ptr;
public:
 Auto_ptr3(T* ptr = nullptr)
  :m_ptr(ptr)
 {
 }

 ~Auto_ptr3()
 {
  delete m_ptr;
 }

 // Copy constructor
 // Do deep copy of a.m_ptr to m_ptr
 Auto_ptr3(const Auto_ptr3& a)
 {
  m_ptr = new T;
  *m_ptr = *a.m_ptr;
 }

 // Copy assignment
 // Do deep copy of a.m_ptr to m_ptr
 Auto_ptr3& operator=(const Auto_ptr3& a)
 {
  // Self-assignment detection
  if (&a == this)
   return *this;

  // Release any resource we're holding
  delete m_ptr;

  // Copy the resource
  m_ptr = new T;
  *m_ptr = *a.m_ptr;

  return *this;
 }

 T& operator*() const { return *m_ptr; }
 T* operator->() const { return m_ptr; }
 bool isNull() const { return m_ptr == nullptr; }
};

class Resource
{
public:
 Resource() { std::cout << "Resource acquired\n"; }
 ~Resource() { std::cout << "Resource destroyed\n"; }
};

Auto_ptr3<Resource> generateResource()
{
 Auto_ptr3<Resource> res(new Resource);
 return res; // this return value will invoke the copy constructor
}

int main()
{
 Auto_ptr3<Resource> mainres;
 mainres = generateResource(); // this assignment will invoke the copy
assignment

 return 0;
}
```

In this program, we're using a function named generateResource() to create a smart pointer encapsulated resource, which is then passed back to function main(). Function main() then assigns that to an existing Auto_ptr3 object.

When this program is run, it prints:

```
Resource acquired
Resource acquired
Resource destroyed
Resource acquired
Resource destroyed
Resource destroyed
```

(Note: You may only get 4 outputs if your compiler elides the return value from function generateResource())

That's a lot of resource creation and destruction going on for such a simple program! What's going on here?

Let's take a closer look. There are 6 key steps that happen in this program (one for each printed message):

1. **Inside generateResource(), local variable res is created and initialized with a dynamically allocated Resource, which causes the first "Resource acquired".**
2. **Res is returned back to main() by value. We return by value here because res is a local variable -- it can't be returned by address or reference because res will be destroyed when generateResource() ends. So res is copy constructed into a temporary object. Since our copy constructor does a deep copy, a new Resource is allocated here, which causes the second "Resource acquired".**
3. **Res goes out of scope, destroying the originally created Resource, which causes the first "Resource destroyed".**
4. **The temporary object is assigned to mainres by copy assignment. Since our copy assignment also does a deep copy, a new Resource is allocated, causing yet another "Resource acquired".**
5. **The assignment expression ends, and the temporary object goes out of expression scope and is destroyed, causing a "Resource destroyed".**
6. **At the end of main(), mainres goes out of scope, and our final "Resource destroyed" is displayed.**

So, in short, because we call the copy constructor once to copy construct res to a temporary, and copy assignment once to copy the temporary into mainres, we end up allocating and destroying 3 separate objects in total.

Inefficient, but at least it doesn't crash!

However, with move semantics, we can do better.

Move constructors and move assignment

C++11 defines two new functions in service of move semantics: a move constructor, and a move assignment operator. Whereas the goal of the copy constructor and copy assignment is to make a copy of one object to another, the goal of the move constructor and move assignment is to move ownership of the resources from one object to another (which is typically much less expensive than making a copy).

Defining a move constructor and move assignment work analogously to their copy counterparts. However, whereas the copy flavors of these functions take a const l-value reference parameter, the move flavors of these functions use non-const r-value reference parameters.

Here's the same Auto_ptr3 class as above, with a move constructor and move assignment operator added. We've left in the deep-copying copy constructor and copy assignment operator for comparison purposes.

```
1  #include <iostream>
2
3  template<class T>
4  class Auto_ptr4
5  {
6   T* m_ptr;
7  public:
8   Auto_ptr4(T* ptr = nullptr)
     :m_ptr(ptr)
9   {
10  }
```

```
 ~Auto_ptr4()
 {
  delete m_ptr;
 }

 // Copy constructor
 // Do deep copy of a.m_ptr to m_ptr
 Auto_ptr4(const Auto_ptr4& a)
 {
  m_ptr = new T;
  *m_ptr = *a.m_ptr;
 }

 // Move constructor
 // Transfer ownership of a.m_ptr to m_ptr
 Auto_ptr4(Auto_ptr4&& a) noexcept
  : m_ptr(a.m_ptr)
 {
  a.m_ptr = nullptr; // we'll talk more about this line below
 }

 // Copy assignment
 // Do deep copy of a.m_ptr to m_ptr
 Auto_ptr4& operator=(const Auto_ptr4& a)
 {
  // Self-assignment detection
  if (&a == this)
   return *this;

  // Release any resource we're holding
  delete m_ptr;

  // Copy the resource
  m_ptr = new T;
  *m_ptr = *a.m_ptr;

  return *this;
 }

 // Move assignment
 // Transfer ownership of a.m_ptr to m_ptr
 Auto_ptr4& operator=(Auto_ptr4&& a) noexcept
 {
  // Self-assignment detection
  if (&a == this)
   return *this;

  // Release any resource we're holding
  delete m_ptr;

  // Transfer ownership of a.m_ptr to m_ptr
  m_ptr = a.m_ptr;
  a.m_ptr = nullptr; // we'll talk more about this line below

  return *this;
 }

 T& operator*() const { return *m_ptr; }
 T* operator->() const { return m_ptr; }
 bool isNull() const { return m_ptr == nullptr; }
};

class Resource
{
public:
 Resource() { std::cout << "Resource acquired\n"; }
 ~Resource() { std::cout << "Resource destroyed\n"; }
};

Auto_ptr4<Resource> generateResource()
{
 Auto_ptr4<Resource> res(new Resource);
 return res; // this return value will invoke the move constructor
}

int main()
{
 Auto_ptr4<Resource> mainres;
 mainres = generateResource(); // this assignment will invoke the move
assignment
```

```
    return 0;
}
```

The move constructor and move assignment operator are simple. Instead of deep copying the source object (a) into the implicit object, we simply move (steal) the source object's resources. This involves shallow copying the source pointer into the implicit object, then setting the source pointer to null.

When run, this program prints:

```
Resource acquired
Resource destroyed
```

That's much better!

The flow of the program is exactly the same as before. However, instead of calling the copy constructor and copy assignment operators, this program calls the move constructor and move assignment operators. Looking a little more deeply:

1. Inside generateResource(), local variable res is created and initialized with a dynamically allocated Resource, which causes the first "Resource acquired".
2. Res is returned back to main() by value. Res is move constructed into a temporary object, transferring the dynamically created object stored in res to the temporary object. We'll talk about why this happens below.
3. Res goes out of scope. Because res no longer manages a pointer (it was moved to the temporary), nothing interesting happens here.
4. The temporary object is move assigned to mainres. This transfers the dynamically created object stored in the temporary to mainres.
5. The assignment expression ends, and the temporary object goes out of expression scope and is destroyed. However, because the temporary no longer manages a pointer (it was moved to mainres), nothing interesting happens here either.
6. At the end of main(), mainres goes out of scope, and our final "Resource destroyed" is displayed.

So instead of copying our Resource twice (once for the copy constructor and once for the copy assignment), we transfer it twice. This is more efficient, as Resource is only constructed and destroyed once instead of three times.

When are the move constructor and move assignment called?

The move constructor and move assignment are called when those functions have been defined, and the argument for construction or assignment is an r-value. Most typically, this r-value will be a literal or temporary value.

In most cases, a move constructor and move assignment operator will not be provided by default, unless the class does not have any defined copy constructors, copy assignment, move assignment, or destructors. However, the default move constructor and move assignment do the same thing as the default copy constructor and copy assignment (make copies, not do moves).

Rule

If you want a move constructor and move assignment that do moves, you'll need to write them yourself.

The key insight behind move semantics

You now have enough context to understand the key insight behind move semantics.

If we construct an object or do an assignment where the argument is an l-value, the only thing we can reasonably do is copy the l-value. We can't assume it's safe to alter the l-value, because it may be used again later in the program. If we have an expression "a = b", we wouldn't reasonably expect b to be changed in any way.

However, if we construct an object or do an assignment where the argument is an r-value, then we know that r-value is just a temporary object of some kind. Instead of copying it (which can be expensive), we can simply transfer its resources (which is cheap) to

the object we're constructing or assigning. This is safe to do because the temporary will be destroyed at the end of the expression anyway, so we know it will never be used again!

C++11, through r-value references, gives us the ability to provide different behaviors when the argument is an r-value vs an l-value, enabling us to make smarter and more efficient decisions about how our objects should behave.

Move functions should always leave both objects in a well-defined state

In the above examples, both the move constructor and move assignment functions set a.m_ptr to nullptr. This may seem extraneous -- after all, if "a" is a temporary r-value, why bother doing "cleanup" if parameter "a" is going to be destroyed anyway?

The answer is simple: When "a" goes out of scope, a's destructor will be called, and a.m_ptr will be deleted. If at that point, a.m_ptr is still pointing to the same object as m_ptr, then m_ptr will be left as a dangling pointer. When the object containing m_ptr eventually gets used (or destroyed), we'll get undefined behavior.

Additionally, in the next lesson we'll see cases where "a" can be an l-value. In such a case, "a" wouldn't be destroyed immediately, and could be queried further before its lifetime ends.

Automatic l-values returned by value may be moved instead of copied

In the generateResource() function of the Auto_ptr4 example above, when variable res is returned by value, it is moved instead of copied, even though res is an l-value. The C++ specification has a special rule that says automatic objects returned from a function by value can be moved even if they are l-values. This makes sense, since res was going to be destroyed at the end of the function anyway! We might as well steal its resources instead of making an expensive and unnecessary copy.

Although the compiler can move l-value return values, in some cases it may be able to do even better by simply eliding the copy altogether (which avoids the need to make a copy or do a move at all). In such a case, neither the copy constructor nor move constructor would be called.

Disabling copying

In the Auto_ptr4 class above, we left in the copy constructor and assignment operator for comparison purposes. But in move-enabled classes, it is sometimes desirable to delete the copy constructor and copy assignment functions to ensure copies aren't made. In the case of our Auto_ptr class, we don't want to copy our templated object T -- both because it's expensive, and whatever class T is may not even support copying!

Here's a version of Auto_ptr that supports move semantics but not copy semantics:

```cpp
#include <iostream>

template<class T>
class Auto_ptr5
{
 T* m_ptr;
public:
 Auto_ptr5(T* ptr = nullptr)
  :m_ptr(ptr)
 {
 }

 ~Auto_ptr5()
 {
  delete m_ptr;
 }

 // Copy constructor -- no copying allowed!
 Auto_ptr5(const Auto_ptr5& a) = delete;

 // Move constructor
 // Transfer ownership of a.m_ptr to m_ptr
 Auto_ptr5(Auto_ptr5&& a) noexcept
  : m_ptr(a.m_ptr)
 {
  a.m_ptr = nullptr;
 }

 // Copy assignment -- no copying allowed!
 Auto_ptr5& operator=(const Auto_ptr5& a) =
delete;

 // Move assignment
 // Transfer ownership of a.m_ptr to m_ptr
 Auto_ptr5& operator=(Auto_ptr5&& a) noexcept
 {
  // Self-assignment detection
  if (&a == this)
   return *this;

  // Release any resource we're holding
  delete m_ptr;

  // Transfer ownership of a.m_ptr to m_ptr
  m_ptr = a.m_ptr;
  a.m_ptr = nullptr;

  return *this;
 }

 T& operator*() const { return *m_ptr; }
 T* operator->() const { return m_ptr; }
 bool isNull() const { return m_ptr == nullptr; }
};
```

If you were to try to pass an Auto_ptr5 l-value to a function by value, the compiler would complain that the copy constructor required to initialize the function argument has been deleted. This is good, because we should probably be passing Auto_ptr5 by const l-value reference anyway!

Auto_ptr5 is (finally) a good smart pointer class. And, in fact the standard library contains a class very much like this one (that you should use instead), named std::unique_ptr. We'll talk more about std::unique_ptr later in this chapter.

Another example

Let's take a look at another class that uses dynamic memory: a simple dynamic templated array. This class contains a deep-copying copy constructor and copy assignment operator.

```cpp
#include <iostream>

template <class T>
class DynamicArray
{
private:
 T* m_array;
 int m_length;

public:
 DynamicArray(int length)
  : m_array(new T[length]), m_length(length)
 {
 }

 ~DynamicArray()
 {
  delete[] m_array;
 }

 // Copy constructor
 DynamicArray(const DynamicArray &arr)
  : m_length(arr.m_length)
 {
  m_array = new T[m_length];
  for (int i = 0; i < m_length; ++i)
   m_array[i] = arr.m_array[i];
 }

 // Copy assignment
 DynamicArray& operator=(const DynamicArray &arr)
 {
  if (&arr == this)
   return *this;

  delete[] m_array;

  m_length = arr.m_length;
  m_array = new T[m_length];

  for (int i = 0; i < m_length; ++i)
   m_array[i] = arr.m_array[i];

  return *this;
 }

 int getLength() const { return m_length; }
 T& operator[](int index) { return m_array[index]; }
 const T& operator[](int index) const { return
m_array[index]; }

};
```

Now let's use this class in a program. To show you how this class performs when we allocate a million integers on the heap, we're going to leverage the Timer class we developed in lesson 12.18 -- Timing your code. We'll use the Timer class to time how fast our code runs, and show you the performance difference between copying and moving.

```
#include <iostream>
#include <chrono> // for std::chrono functions

// Uses the above DynamicArray class

class Timer
{
private:
 // Type aliases to make accessing nested type easier
 using clock_t = std::chrono::high_resolution_clock;
 using second_t = std::chrono::duration<double, std::ratio<1> >;

 std::chrono::time_point<clock_t> m_beg;

public:
 Timer() : m_beg(clock_t::now())
 {
 }

 void reset()
 {
  m_beg = clock_t::now();
 }

 double elapsed() const
 {
  return std::chrono::duration_cast<second_t>(clock_t::now() -
m_beg).count();
 }
};

// Return a copy of arr with all of the values doubled
DynamicArray<int> cloneArrayAndDouble(const DynamicArray<int> &arr)
{
 DynamicArray<int> dbl(arr.getLength());
 for (int i = 0; i < arr.getLength(); ++i)
  dbl[i] = arr[i] * 2;

 return dbl;
}

int main()
{
 Timer t;

 DynamicArray<int> arr(1000000);

 for (int i = 0; i < arr.getLength(); i++)
  arr[i] = i;

 arr = cloneArrayAndDouble(arr);

 std::cout << t.elapsed();
}
```

On one of the author's machines, in release mode, this program executed in 0.00825559 seconds.

Now let's run the same program again, replacing the copy constructor and copy assignment with a move constructor and move assignment.

```
template <class T>
class DynamicArray
{
private:
 T* m_array;
 int m_length;

public:
 DynamicArray(int length)
  : m_array(new T[length]), m_length(length)
 {
 }

 ~DynamicArray()
 {
  delete[] m_array;
 }

 // Copy constructor
```

```
 // Copy constructor
 DynamicArray(const DynamicArray &arr) = delete;

 // Copy assignment
 DynamicArray& operator=(const DynamicArray &arr) = delete;

 // Move constructor
 DynamicArray(DynamicArray &&arr) noexcept
  : m_length(arr.m_length), m_array(arr.m_array)
 {
  arr.m_length = 0;
  arr.m_array = nullptr;
 }

 // Move assignment
 DynamicArray& operator=(DynamicArray &&arr) noexcept
 {
  if (&arr == this)
   return *this;

  delete[] m_array;

  m_length = arr.m_length;
  m_array = arr.m_array;
  arr.m_length = 0;
  arr.m_array = nullptr;

  return *this;
 }

 int getLength() const { return m_length; }
 T& operator[](int index) { return m_array[index]; }
 const T& operator[](int index) const { return m_array[index]; }

};

#include <iostream>
#include <chrono> // for std::chrono functions

class Timer
{
private:
 // Type aliases to make accessing nested type easier
 using clock_t = std::chrono::high_resolution_clock;
 using second_t = std::chrono::duration<double, std::ratio<1> >;

 std::chrono::time_point<clock_t> m_beg;

public:
 Timer() : m_beg(clock_t::now())
 {
 }

 void reset()
 {
  m_beg = clock_t::now();
 }

 double elapsed() const
 {
  return std::chrono::duration_cast<second_t>(clock_t::now() -
m_beg).count();
 }
};

// Return a copy of arr with all of the values doubled
DynamicArray<int> cloneArrayAndDouble(const DynamicArray<int> &arr)
{
 DynamicArray<int> dbl(arr.getLength());
 for (int i = 0; i < arr.getLength(); ++i)
  dbl[i] = arr[i] * 2;

 return dbl;
}

int main()
{
 Timer t;

 DynamicArray<int> arr(1000000);

 for (int i = 0; i < arr.getLength(); i++)
```

```
    arr[i] = i;

  arr = cloneArrayAndDouble(arr);

  std::cout << t.elapsed();
}
```

On the same machine, this program executed in 0.0056 seconds.

Comparing the runtime of the two programs, 0.0056 / 0.00825559 = 67.8%. The move version was almost 33% faster!

Next lesson

M.4 std::move

Back to table of contents

Previous lesson

M.2 R-value references

204 COMMENTS

Newest