A New Asynchronous Method Invocation API for Ice for Java

Michi Henning, Chief Scientist, ZeroC, Inc.

Introduction

Ice has had an API for asynchronous method invocation (AMI) since its inception. However, that API is quite verbose and not all that easy to use, as well as inflexible. With the release of Ice 3.4, ZeroC introduces a new API for asynchronous method invocation that does not suffer from these problems and provides programmers with far more choice as to how they can structure their code. The new API is available for Java, .NET, C++, and Python.

This article provides an overview of the new Java API and explains its most important features. (As always, you should consult the Ice Manual for complete documentation.) Companion articles describe the corresponding new APIs for .NET, C++, and Python.

The old API is still available and it is possible to use both the new and the old API in the same program. This allows you to gradually migrate code away from the old API without having to change all your code at once. For new code, we strongly recommend that you use the new API. (The old API is deprecated and will eventually be removed entirely.)

Contents

The Problems of the Old API

Suppose you have the following Slice definition:

```
// Slice
module Demo
{
     ["ami"]
     interface Employees
     {
         string getName(int number);
     };
};
```

With the old API, to allow you to invoke `getName` asynchronously, **`slice2java`** generates the following proxy methods:

```
public interface EmployeesPrx extends Ice.ObjectPrx
{
    public boolean getName_async(AMI_Employees_getName __cb,
                                  int number);
    public boolean getName_async(
                             AMI_Employees_getName __cb,
                             int number,
                             java.util.Map<String, String> __ctx);
}
```

Note that the `getName_async` method is overloaded so you can add a per-invocation context instead of sending the default context. (The new API also permits you to supply a context, but we do not discuss contexts further in this article. Please see the Ice Manual for a full description.)

To call `getName` asynchronously, you must implement a class that derives from the `Demo.AMI_Employees_getName` class that is generated by **`slice2java`**. Your derived class must implement two methods, `ice_response` and `ice_exception`. (If an operation raises user exceptions, you need a third method to that is called for user exceptions.) These methods are the callback methods that are called by the Ice run time to inform you when an asynchronous call of `getName` completes. For example:

```
private class AMI_Employees_getNameI
     extends Demo.AMI_Employees_getName
{
     public void ice_response(String name)
     {
         System.out.println("Name is: " + name);
     }

     public void ice_exception(Ice.LocalException ex)
     {
         System.out.println("Exception is: " + ex);
     }
}
```

When you call `getName` asynchronously, you must supply an instance of this callback class, for example:

```
EmployeesPrx e = ...;
AMI_Employees_getNameI cb = new AMI_Employees_getNameI();
e.getName_async(cb, 99);
```

This invokes the `getName` operation in the server without blocking the caller; some time later, once the call completes, the Ice run time calls `ice_response` on the callback instance (if the call succeeded) or `ice_exception` (if the call failed).

This is fine as far as it goes. However, there are a number of disadvantages with this API:

- You must implement a separate class for every operation you call asynchronously.
- The class must implement two methods with the names `ice_response` and `ice_exception`.
- There is no way to poll for completion of an asynchronous call.
- There is no simple way to block until a particular call completes. (To achieve this, you have to use a monitor that is signaled from the `ice_response` and `ice_exception` callbacks.)
- The base class from which you derive your callback class stores the state of the asynchronous call. This means that you cannot use the same callback instance for multiple concurrent calls. (You can reuse a callback instance, but only once the previous invocation using that instance is complete.)

The above points boil down to two major points: verbosity and inflexibility. The verbosity is not immediately obvious, but becomes clear when you consider an application of more realistic size. For example, if you have eight interfaces with six operations each (all of which you want to call asynchronously), you must implement 48 classes and 96 methods. This is tedious, to say the least.

The inflexibility of the API also is a concern: you cannot poll for call completion and you cannot easily block until a particular call is complete. The *only* way to get the results of an invocation is an asynchronous callback. Moreover, there is no easy way to share code on a call-by-call basis. For example, if in a particular section of your application, you want to treat a particular error condition in a different way, you must write a new callback class for every operation that implements the different error handling (or, alternatively, pass additional state into your callback instance so it can select inside the `ice_exception` callback which behavior is needed).

In summary, the old API is as rigid as steel: there is one and only one way of doing things. This is inconvenient if, for example, you have many simple get/set operations that substantially perform the same actions when they complete, or if you would like to block until a particular invocation or group of invocations is complete.

The New Approach

The new asynchronous API eliminates both the verbosity and inflexibility of the old API. Not only can you do more things with the new API, but you can also do so with less code. In turn, this reduces development time and maintenance effort.

The following sections provide an overview of the capabilities of the new API and show some examples of how and why you might want to use a particular API feature.

No Need for Metadata

Here is our simple `Employees` interface once more:

```
// Slice
module Demo
{
    interface Employees
    {
        string getName(int number);
    };
};
```

Note that the `["ami"]` metadata directive is absent. This is because the new asynchronous API is always generated by **`slice2java`**, so no separate metadata directive is necessary. (If you do add the `["ami"]` directive, **`slice2java`** generates both the old and the new API; you can use both APIs in the same program.)

Basic Asynchronous Invocations

The asynchronous proxy methods look as follows:

```
public interface EmployeesPrx extends Ice.ObjectPrx
{
    public Ice.AsyncResult begin_getName(int number);
    public String end_getName(Ice.AsyncResult __result);
}
```

Note that the `getName` operation now has a `begin_` method and an `end_` method. The `begin_getName` method starts the asynchronous call and is guaranteed not to block the caller. The `end_getName` method is used to collect the result of the invocation. If, at the time the client calls `end_getName`, the operation is not complete yet, `end_getName` blocks the calling thread until the call is complete. On the other hand, if the call completed earlier, some time after the client called `begin_getName` but before it calls `end_getName`, `end_getName` completes immediately.

The API generated by **`slice2java`** has several overloads of the `begin_` method (of which we only show the first one in the preceding example); the other overloads deal with contexts (see the Ice Manual for details) as well as callbacks, which we will discuss shortly.

Here is an example of how a client can make such an asynchronous invocation:

```
EmployeesPrx e = ...;
Ice.AsyncResult r = e.begin_getName(99); // Does not block

// Continue to do other things here...

String name = e.end_getName(r); // Blocks until result is available
```


Now, at first glance, this does not look particularly useful: if you call `end_getName` immediately after calling `begin_getName`, you get the same effect as having used an ordinary synchronous call.

However, because `begin_getName` is guaranteed not to block, you can continue to do other things. This is useful if you know that a particular operation invocation may take some time, and there are other tasks you can perform before you need to collect the result of the invocation. That way, you get increased concurrency between client and server without having to use separate threads.

The `begin_` method has one parameter for each in-parameter of the corresponding Slice operation. Similarly, the `end_` method has one parameter for each out-parameter of the corresponding Slice operation. (If a Slice operation has a return value, the `end_` method returns that value in the same way that a synchronous invocation would.)

If an operation raises a user exception or Ice run-time exception, the exception is thrown by the `end_` method. (The `begin_` method does not throw Ice exceptions other than `CommunicatorDestroyedException`.)

Polling and Waiting for Call Completion

Note that the `begin_` method returns an `AsyncResult` instance. This instance encapsulates the state of the asynchronous call and allows you to learn something about the details of the call. You must pass the `AsyncResult` you obtained from the `begin_` method to the corresponding `end_` method. The information in the `AsyncResult` allows the Ice run time to locate the reply from the server and to call the appropriate unmarshaling code. (The unmarshaling of the reply is done by the `end_` method, that is, the Ice run time stores the reply from the server until the `end_` method is called, and the decoding of the reply is done by the `end_` method.)

The `AsyncResult` class contains a few methods that allow you to check call progress and completion:

```
public class AsyncResult
{
    public final boolean sentSynchronously();

    public final boolean isSent();
    public final void waitForSent();

    public final boolean isCompleted();
    public final void waitForCompleted();

    // ...
}
```

`sentSynchronously` reports whether the Ice run time was able to immediately pass the invocation to the client's local transport. The method returns true if the invocation was written to the local transport immediately; otherwise, the method returns false, indicating that the Ice run time queued the invocation for later transmission because the local transport could not accept it at the time the `begin_` method was called.

ZeroC

isSent reports whether, at that time, the invocation has been passed to the client's local transport (whether it was initially queued for transmission or not). waitForSent blocks the calling thread until the local transport has accepted the invocation and returns immediately if the invocation was written to the local transport earlier.

The isCompleted method returns true if the Ice run time has received the server's reply for the invocation (whether successful or indicating an exception) and false, otherwise. The waitForCompleted method blocks the calling thread until the reply from the server has been received and returns immediately if the reply was received earlier.

One way to use these methods is to achieve better concurrency between client and server for data transfers. For example, suppose we have an interface that permits the client to send a file to the server. Because files are often larger than what can be transmitted with a single remote procedure call, the interface allows the client to send the file in chunks:

```
// Slice
module Demo
{
    interface FileTransfer
    {
        void send(int offset, ByteSeq bytes);
    };
};
```

The client can repeatedly call send to send a chunk of the file, indicating at which offset in the file the chunk belongs. A naïve way to transmit a file would be along the following lines:

```
FileHandle file = open(...);
FileTransferPrx ft = ...;
int chunkSize = ...;
int offset = 0;
while(!file.eof())
{
    byte[] bs;
    bs = file.read(chunkSize); // Read a chunk
    ft.send(offset, bs);       // Send the chunk
    offset += bs.length;
}
```

This works, but not very well: because the client makes a synchronous call, it writes each chunk on the wire and then waits for the server to receive the data, process it, and return a reply before writing the next chunk. This means that both client and server spend much of their time doing nothing—the client does nothing while the server processes the data, and the server does nothing while it waits for the client to send the next chunk.

Using asynchronous calls, we can improve on this considerably:

```
FileHandle file = open(...);
FileTransferPrx ft = ...;
int chunkSize = ...;
int offset = 0;
```

```
LinkedList<Ice.AsyncResult> results =
    new LinkedList<Ice.AsyncResult>();
int numRequests = 5;

while(!file.eof())
{
    byte[] bs;
    bs = file.read(chunkSize);

    // Send up to numRequests + 1 chunks asynchronously.
    Ice.AsyncResult r = ft.begin_send(offset, bs);
    offset += bs.length;

    // Wait until this request has been passed to the transport.
    r.waitForSent();
    results.add(r);

    // Once there are more than numRequests, wait for the least
    // recent one to complete.
    while(results.size() > numRequests)
    {
        Ice.AsyncResult r = results.getFirst();
        results.removeFirst();
        r.waitForCompleted();
    }
}

// Wait for any remaining requests to complete.
while(results.size() > 0)
{
    Ice.AsyncResult r = results.getFirst();
    results.removeFirst;
    r.waitForCompleted();
}
```

With this code, the client sends up to `numRequests + 1` chunks before it waits for the least recent one of those requests to complete. In other words, the client sends the next request without waiting for the preceding request to complete, up to the limit set by `numRequests`. In effect, this allows the client to "keep the pipe to the server full of data": the client keeps sending data, so both client and server continuously do work.

Obviously, the correct chunk size and value of `numRequests` depends on the bandwidth of the network as well as the amount of time taken by the server to process each request. However, with a little testing, you can quickly zoom in on the point where making the requests larger or queuing more requests no longer improves performance. With this technique, you can realize the full bandwidth of the link to within a percent or two of the theoretical bandwidth limit of a native socket connection.

Completion Callbacks

The `begin_` method is overloaded to allow you to supply a callback object with a `completed` method that is called by the Ice run time when an asynchronous invocation completes. You must implement a class that derives from `Ice.AsyncCallback` that implements this method. The implementation of `completed` is expected to call the `end_` method for the corresponding invocation. For example, you could write the callback class for an invocation of `getName` as follows:

```
public class MyCallback extends Ice.AsyncCallback
{
    public void completed(Ice.AsyncResult r)
    {
        EmployeesPrx e = (EmployeesPrx)r.getProxy();
        try
        {
            String name = e.end_getName(r);
            System.out.println("Name is: " + name);
        }
        catch(Ice.LocalException ex)
        {
            System.err.println("Exception is: " + ex);
        }
    }
}
```

Note the `getProxy` method on the `AsyncResult` that is passed to `completed`: it returns the proxy that was used to invoke the `begin_` method. The return value of `getProxy` is of type `Ice.ObjectPrx`, so we need to down-cast it to an `EmployeePrx` before we can invoke `end_getName`.

Having written the completion callback, the question is how we inform the Ice run time that we want it to call `completed` when a `getName` call completes. We achieve this by passing an instance of the callback class to the `begin_` method:

```
EmployeesPrx e = ...;

MyCallback cb = new MyCallback();
e.begin_getName(99, cb);
```

Some time after calling `begin_getName` in this way, the Ice run time calls `completed` on your callback instance and, in turn, `completed` calls the `end_` method to unmarshal the server's reply.

This scheme is simpler than the old asynchronous API because you only need a single callback method instead of two.

A more terse way to achieve the same thing is to use an anonymous class:

```
EmployeesPrx e = ...;

e.begin_getName(99,
                new Ice.AsyncCallback()
                {
                    public void completed(Ice.AsyncResult r)
                    {
                        EmployeesPrx e = (EmployeesPrx)r.getProxy();
                        try
                        {
                            String name = e.end_getName(r);
```


ZeroC

```
                    System.out.println("Name is: " + name);
                }
                catch(Ice.LocalException ex)
                {
                    System.err.println("Exception: " + ex);
                }
            }
        });
```

This style is useful particularly for callbacks that do only a small amount of work because the code that starts the call and the code that processes the results are physically close together.

Using a Single Callback Method for Many Operations

The approach we have explored so far all requires you to implement a separate callback class for each operation. However, you can easily create a generic callback method that can handle the results for many different operations.

Consider a modified version of our `Employees` interface:

```
// Slice
module Demo
{
    interface Employees
    {
        string getName(int number);
        int getNumber(string name);
    };
};
```

This is the same interface as previously, but with an additional `getNumber` operation. We can now create a single callback class that can deal with results from invoking either operation, as well as easily share common exception handling:

The actions taken by the callback depend on the operation whose results are being processed: it is the operation that determines the type of the return value and out-parameters (if any). Suppose we want to use a single callback method to process the result of both the `getName` and `getNumber` operations on our `Employees` interface. One way to achieve this is to retrieve the operation name from the `AsyncResult` and use it to decide which `end_` method to call:

```
public class MyCallback extends Ice.AsyncCallback
{
    public void completed(Ice.AsyncResult r)
    {
        EmployeesPrx e = (EmployeesPrx)r.getProxy();
        try
        {
            String op = r.getOperation();
            if(op.equals(("getName"))
            {
                String name = e.end_getName(r);
                System.out.println("Name is: " + name);
            }
            else
```

```
                {
                    int number = e.end_getNumber(r);
                    System.out.println("Number is: " + name);
                }
            }
            catch(Ice.LocalException ex)
            {
                System.err.println("Exception is: " + ex);
            }
        }
}
```

With this implementation, the calling end can use the same callback method for both operations:

```
MyCallback cb = new MyCallback();

e.begin_getName(99, cb);
e.begin_getNumber("Fred", cb);
```

Another implementation that achieves the same thing is to store a map in the callback class that maps the operation name to an enumerator. The callback method uses the operation name to efficiently retrieve the enumerator (instead of performing a series of string comparisons) and, with a `switch` statement, to select what `end_` method to call and how to process the operation result.

Which implementation technique you end up using depends entirely on your application. The point is that the API is flexible and therefore can adapt to your needs. This is particularly useful if you already have a large body of code and want to call that code when the results of an invocation are ready: you can easily choose an implementation that matches your needs instead of having to use the rigid one-size-fits-all approach of the old API.

Passing State from the `begin_` Method to the `end_` Method

It is common for applications to have some shared state that is established when an asynchronous call is started, and needed again when that call completes. As an example, consider an application that asynchronously starts a number of operations and, as each operation completes, needs to update different user interface elements with the results. In this case, the code calling the `begin_` method knows which user interface element should receive the update, and the code inside `end_` method needs access to that element.

Assuming that we have a `Widget` class that designates a particular user interface element, you could pass different widgets by storing the widget to be used as a member of your callback class:

```
EmployeesPrx e = ...;
Widget widget1 = ...;
Widget widget2 = ...;

public class MyCallback extends Ice.AsyncCallback
{
    public MyCallback(Widget w)
    {
        _w = w;
    }
```

```
    private Widget _w;

    public void completed(Ice.AsyncResult r)
    {
        EmployeesPrx e = (EmployeesPrx)r.getProxy();
        try
        {
            String name = e.end_getName(r);
            _w.writeString(name);
        }
        catch(Ice.LocalException ex)
        {
            System.err.println("Exception is: " + ex);
        }
    }
}

// Invoke the getName operation with different widget callbacks.
e.begin_getName(99, new MyCallback(widget1));
e.begin_getName(24, new MyCallback(widget2));
```

This example assumes that widgets have a `writeString` method that updates the relevant UI element. The callback class provides a simple and effective way for you to pass state between the point where an operation is invoked and the point where its results are processed. Moreover, if you have a number of operations that share common state, you can pass the same callback instance to multiple invocations. (If you do this, your callback methods may need to use synchronization.) This allows you to tailor your classes to match the semantics of your application.

Type-Safe Callbacks

One thing you may have noticed with the API we have explored so far is that it is not entirely type-safe:

- You must down-cast the return value of `getProxy` to the correct proxy type before you can call the `end_` method.
- You must call the correct `end_` method to match the operation called by the `begin_` method.
- You must remember to catch exceptions when you call the `end_` method; if you forget to do this, you will not know that the operation failed. (If a callback method throws an exception, the Ice run time ignores the exception after logging a warning.)

This lack of type-safety is the price we pay for the generic nature of the API. If you do not require the flexibility of the generic API, you can instead use a type-safe API. That API is somewhat less flexible but, in return, does not require you to perform any down-casts, select the correct `end_` method, or catch exceptions.

To use type-safe callbacks, you implement a callback class that has one method that receives the operation result if the operation succeeds, and another method that receives the exception that is raised if the operation fails. The callback class must extend a Slice-generated base class. For example:

```
public class MyCallback extends Demo.Callback_Employees_getName
{
```

```
    public void response(String name)
    {
        System.out.println("Name is: " + name);
    }

    public void exception(Ice.LocalException ex)
    {
        System.err.println("Exception is: " + ex);
    }
}
```

The callback methods must have the names `response` and `exception`. The Ice run time calls `response` if the operation succeeds and `exception` if the operation raises an exception. The name of the base class is `Callback_`*`<interface>`*`_`*`<operation>`*.

Note that the callback methods are now strongly typed: the return value of the operation becomes an in-parameter of the correct type. (If an operation has out-parameters, these become additional in-parameters for the `response` method.)

This style of callback is similar to the old API with its `ice_response` and `ice_exception` methods. However, it allows you to use the same callback instance for multiple concurrent invocations because the callback instance does not store any state that is needed by the Ice run time.

At the calling end, you call the `begin_` method as you would for the generic API:

```
MyCallback cb = new MyCallback();

e.begin_getName(99, cb);
```

If you want to pass state from the calling end to the callback, you can use the same approach as for the generic API.

Flow Control

Asynchronous method invocations never block the thread that calls the `begin_` method: the Ice run time checks to see whether it can write the request to the local transport. If it can, it does so immediately in the caller's thread. (In that case, `AsyncResult.sentSynchronously` returns true.) Alternatively, if the local transport does not have sufficient buffer space to accept the request, the Ice run time queues the request internally for later transmission in the background. (In that case, `AsyncResult.sentSynchronously` returns false.)

This creates a potential problem: if a client sends many asynchronous requests at the time the server is too busy to keep up with them, the requests pile up in the client-side run time until, eventually, the client runs out of memory.

The API provides a way for you to implement flow control by counting the number of requests that are queued so, if that number exceeds some threshold, the client stops invoking more operations until some of the queued operations have drained out of the local transport.

For the generic API, you can create an additional `sent` method:

```
public class MyCallback extends Ice.AsyncCallback
{
    public void completed(Ice.AsyncResult r)
    {
        // ...
    }

    public void sent(Ice.AsyncResult r)
    {
        // ...
    }
}
```

You inform the Ice run time that you want to be informed when a call has been passed to the local transport by passing the callback instance as usual:

```
e.begin_getName(99, new MyCallback());
```

If the Ice run time can immediately pass the request to the local transport, it does so and invokes the `sent` method from the thread that calls the `begin_` method. On the other hand, if the run time has to queue the request, it calls the `sent` method from a different thread once it has written the request to the local transport. In addition, you can find out from the `AsyncResult` that is returned by the `begin_` method whether the request was sent synchronously or was queued, by calling `sentSynchronously`.

For the type-safe API, the sent callback has the following signature:

```
void sent(boolean sentSynchronously);
```

The `sentSynchronously` parameter is true if the invocation was sent immediately, and false if it was queued and sent at a later time.

The `sent` methods allow you to limit the number of queued requests by counting the number of requests that are queued and decrementing the count when the Ice run time passes a request to the local transport. (See the Ice Manual for more detail.)

Oneway Invocations

The new API permits you to invoke operations via oneway proxies just like you would invoke them via twoway proxies. Because oneway operations do not return a result from the server, the type-safe version of the API does not use a `response` callback (only an `exception` callback, in case the operation raised an exception "on the way out", that is, in the client-side run time).

Asynchronous Invocation of Operations on `Ice.Object`

The old API did not permit you to invoke remote operations on `Ice.Object` (such as `ice_ping`) asynchronously. The new API rectifies this and permits you to invoke these operations asynchronously.

Concurrency

The Ice run time always invokes your callback methods from a separate thread, with one exception: it calls the `sent` callback from the thread calling the `begin_` method if the request could be sent synchronously. For the `sent` callback, you know which thread is calling the callback by looking at the `sentSynchronously` member or parameter, so you can take appropriate action to avoid a deadlock.

Summary

The new asynchronous API for Ice is a big improvement over its earlier incarnation. Not only does it provide more features, but it also is less verbose and, due to its flexibility, makes it easier to match the style of interaction with the Ice run time to the needs of your application.

ZeroC will continue to provide the old API for some time. However, be aware that, as of Ice 3.4, the old API is deprecated and that it will eventually be removed. You should therefore use the new API when developing new code, and migrate old code to use the new API.