Microprofile Fault Tolerance

Emily Jiang

1.0, 2017-09-13

Table of Contents

Specification: Microprofile Fault Tolerance

Version: 1.0

Status: Specification

Release: 2017-09-13

Chapter 1. Architecture

This specification defines an easy to use and flexible system for building resilient applications.

1.1. Rational

It is increasingly important to build fault tolerant microservices. Fault tolerance is about leveraging different strategies to guide the execution and result of some logic. Retry policies, bulkheads, and circuit breakers are popular concepts in this area. They dictate whether and when executions should take place, and fallbacks offer an alternative result when an execution does not complete successfully.

As mentioned above, the Fault Tolerance specification is to focus on the following aspects:

- `Timeout`: Define a duration for timeout
- `Retry`: Define a criteria on when to retry
- `Fallback`: provide an alternative solution for a failed execution.
- `CircuitBreaker`: offer a way of fail fast by automatically failing execution to prevent the system overloading and indefinite wait or timeout by the clients.
- `Bulkhead`: isolate failures in part of the system while the rest part of the system can still function.

The main design is to separate execution logic from execution. The execution can be configured with fault tolerance policies, such as RetryPolicy, fallback, Bulkhead and CircuitBreaker.

Hystrix and Failsafe are two popular libraries for handling failures. This specification is to define a standard API and approach for applications to follow in order to achieve the fault tolerance.

This specification introduces the following interceptor bindings:

- `Timeout`
- `Retry`
- `Fallback`
- `CircuitBreaker`
- `Bulkhead`
- `Asynchronous`

Refer to Interceptor Specification for more information on interceptor bindings.

Chapter 2. Relationship to other specifications

This specification defines a set of annotations to be used by classes or methods. The annotations are interceptor bindings. Therefore, this specification depends on the Java Interceptors and Contexts and Dependency Injection specifications define in Java EE platform.

2.1. Relationship to Contexts and Dependency Injection

The Contexts and Dependency Injection (CDI) specification defines a powerful component model to enable loosely coupled architecture design. This specification explores the rich SPI provided by CDI to register an interceptor so that the Fault Tolerance policies can be applied to the method invocation.

2.2. Relationship to Java Interceptors

The Java Interceptors specification defines the basic programming model and semantics for interceptors. This specification uses the typesafe interceptor bindings. The annotations `@Asynchronous`, `@Bulkhead`, `@CircuitBreaker`, `@Fallback`, `@Retry` and `@Timeout` are all interceptor bindings.

These annotations may be bound at the class level or method level. The annotations adhere to the interceptor binding rules defined by Java Interceptors specification.

For instance, if the annotation is bound to the class level, it applies to all business methods of the class. If the component class declares or inherits a class level interceptor binding, it must not be declared final, or have any static, private, or final methods. If a non-static, non-private method of a component class declares a method level interceptor binding, neither the method nor the component class may be declared final.

Since this specification depends on CDI and interceptors specifications, fault tolerance operations have the following restrictions:

- Fault tolerance interceptors bindings must applied on a bean class or bean class method otherwise it is ignored,
- invocation must be business method invocation as defined in CDI specification.
- if a method and its containing class don't have any fault tolerance interceptor binding, it won't be considered as a fault tolerance operation.

2.3. Relationship to MicroProfile Config

The MicroProfile config specification defines a flexible config model to enable microservice configurable and achieve the strict separation of config from code. All parameters on the annotations/interceptor bindings are config properties. They can be configured externally either

via other predefined config sources (e.g. environment variables, system properties or other sources). For an instance, the `maxRetries` parameter on the `@Retry` annotation is a configuration property. It can be configured externally.

Chapter 3. Execution

Use interceptor and annotation to specify the execution and policy configuration. An annotation of Asynchronous has to be specified for any asynchronous calls. Otherwise, synchronous execution is assumed.

Chapter 4. Asynchronous

`Asynchronous` means the execution of the client request will be on a separate thread. This thread should have the correct security context or naming context associated with.

4.1. Asynchronous Usage

A method or a class can be annotated with `@Asynchronous`, which means the method or the methods under the class will be invoked by a separate thread. The method annotated with `@Asynchronous` must return a `Future`, otherwise, `FaultToleranceDefinitionException` occurs.

```
@Asynchronous
public Future<Connection> serviceA() {
   Connection conn = null;
   counterForInvokingServiceA++;
   conn = connectionService();
   return CompletableFuture.completedFuture(conn);
}
```

The above code-snippet means the method serviceA applies the `Asynchronous` policy, which means the invocation will be done by a different thread.

The `@Asynchronous` annotation can be used together with `@Timeout`, `@Fallback`, `@Bulkhead` and `@Retry`. Method invocation will occur in a different thread.

Chapter 5. Timeout

`Timeout` prevents from the execution from waiting forever. It is recommended that a microservice invocation should have timeout associated with.

5.1. Timeout Usage

A method or a class can be annotated with `@Timeout`, which means the method or the methods under the class will have Timeout policy applied.

```
@Timeout(400) // timeout is 400ms
public Connection serviceA() {
   Connection conn = null;
   counterForInvokingServiceA++;
   conn = connectionService();
   return conn;
}
```

The above code-snippet means the method serviceA applies the `Timeout` policy, which is to fail the execution if the execution takes more than 400ms to complete even if it successfully returns.

When a timeout occurs, A `TimeoutException` must be thrown. The `@Timeout` annotation can be used together with `@Fallback`, `@CircuitBreaker`, `@Asynchronous`, `@Bulkhead` and `@Retry`.

When `@Timeout` is used without `@Asynchronous`, the current thread will be interrupted with a call to Thread.interrupt() on reaching the specified timeout duration. The interruption will only work in certain scenarios. The interruption will not work for the following situations:

- The thread is blocked on blocking I/O (database, file read/write), an exception is thrown only in case of waiting for a NIO channel
- The thread isn't waiting (CPU intensive task) and isn't checking for being interrupted
- The thread will catch the interrupted exception (with a general catch block) and will just continue processing, ignoring the interrupt

In the above situations, it is impossible to suspend the execution. The execution thread will finish its process. If the execution takes longer than the specified timeout, the `TimeoutException` will be thrown and the execution result will be discarded.

A `@Fallback` can be specified and it will be invoked if the `TimeoutException` is thrown. If `@Timeout` is used together with `@Retry`, the `TimoutException` will trigger the retry. When `@Timeout` is used with `@CircuitBreaker` and if a `TimeoutException` occurs, the failure will contribute towards the circuit open.

Chapter 6. Retry Policy

In order to recover from a brief network glitch, `@Retry` can be used to invoke the same operation again. The `Retry` policy allows to configure :

- `maxRetries`: the maximum retries
- `delay`: delays between each retry
- `delayUnit`: the delay unit
- `maxDuration`: maximum duration to perform the retry for.
- `durationUnit`: duration unit
- `jitter`: the random vary of retry delays
- `jitterDelayUnit`: the jitter unit
- `retryOn`: specify the failures to retry on
- `abortOn`: specify the failures to abort on

6.1. Retry usage

`@Retry` can be applied to the class or method level. If applied to a class, it means the all methods in the class will have the `@Retry` policy applied. If applied to a method, it means that method will have `@Retry` policy applied. If the `@Retry` policy applied on a class level and on a method level within that class, the method level `@Retry` will override the class-level `@Retry` policy for that particular method.

```
    /**
     * The configured the max retries is 90 but the max duration is 1000ms.
     * Once the duration is reached, no more retries should be performed,
     * even through it has not reached the max retries.
     */
    @Retry(maxRetries = 90, maxDuration= 1000)
    public void serviceB() {
        writingService();
    }

    /**
    * There should be 0-800ms (jitter is -400ms - 400ms) delays
    * between each invocation.
    * there should be at least 4 retries but no more than 10 retries.
    */
    @Retry(delay = 400, maxDuration= 3200, jitter= 400, maxRetries = 10)
    public Connection serviceA() {
        return connectionService();
    }

    /**
    * Sets retry condition, which means Retry will be performed on
    * IOException.
    */
    @Retry(retryOn = {IOException.class})
    public void serviceB() {
        writingService();
    }
```

The `@Retry` annotation can be used together with `@Fallback`, `@CircuitBreaker`, `@Asynchronous`, `@Bulkhead` and `@Timeout`. A `@Fallback` can be specified and it will be invoked if the `@Retry` still fails. If `@Retry` is used together with `@Timeout`, the `TimoutException` will trigger the retry.

Chapter 7. Fallback

Fallbacks are invoked once a Retry or CircuitBreaker has failed enough times.

For a Retry, Fallback is handled any time the Retry would exceed its maximum number of attempts.

For a CircuitBreaker, it is invoked any time the method invocation fails. When the Circuit is open, the Fallback is always invoked.

7.1. Fallback usage

A method or a class can be annotated with `@Fallback`, which means the method or the methods under the class will have Fallback policy applied. There are two ways to specify fallback:

- Specify a FallbackHandler class
- Specify the fallbackMethod

7.1.1. Specify a FallbackHandler class

If a FallbackHandler is registered for a method returning a different type than the FallbackHandler would return, then the container should treat as an error and deployment fails.

FallbackHandlers are meant to be CDI managed, and should follow the life cycle of the scope of the bean.

```
    @Retry(maxRetries = 1)
    @Fallback(StringFallbackHandler.class)
    public String serviceA() {
        counterForInvokingServiceA++;
        return nameService();
    }
```

The above code snippet means when the method failed and retry reaches its maximum retry, the fallback operation will be performed. The method `StringFallbackHandler.handle(ExecutionContext context)` will be invoked. The return type of `StringFallbackHandler.handle(ExecutionContext context)` must be `String`. Otherwise, the `FaultToleranceDefinitionException` exception will be thrown.

7.1.2. Specify the fallbackMethod

This is used if the `fallbackMethod` is on the same class as the method to call fall back.

```
    @Retry(maxRetries = 2)
    @Fallback(fallbackMethod= "fallbackForServiceB")
    public String serviceB() {
        counterForInvokingServiceB++;
       return nameService();
    }

    private String fallbackForServiceB() {
        return "myFallback";
    }
```

The above code snippet means when the method failed and retry reaches its maximum retry, the fallback operation will be performed. The method `fallbackForServiceB` will be invoked. The return type of `fallbackForServiceB` must be `String` and the argument list for `fallbackForServiceB` must be the same as `ServiceB`. Otherwise, the `FaultToleranceDefinitionException` exception will be thrown.

The parameter `value` and `fallbackMethod` on `@Fallback` cannot be specified at the same time. Otherwise, the `FaultToleranceDefinitionException` exception will be thrown.

The fallback should be triggered when an exception occurs. For instance, `BulkheadException`, `CircuitBreakerOpenException`, `TimeoutException` should trigger the fallback.

Chapter 8. Circuit Breaker

Circuit Breaker prevents repeating timeout, so that invoking dysfunctional services or APIs fail fast. There are three circuit states:

- Closed: the service functions well. If a failure occurs, the circuit breaker keeps the record. After the specified failures are reached, the circuit will be opened.
- Open: calls to the circuit breaker fail immediately. After the specified timeout is reached, the circuit transitions to half open state.
- Half-open: one call to the circuit is allowed. If it fails, the circuit will remain open. Otherwise, subsequent calls are allowed. After the specified successful execution, the circuit will be closed.

8.1. Circuit Breaker Usage

A method or a class can be annotated with `@CircuitBreaker`, which means the method or the methods under the class will have CircuitBreaker policy applied.

```
@CircuitBreaker(successThreshold = 10, requestVolumeThreshold = 4, failureRatio=0.75,
delay = 1000)
public Connection serviceA() {
   Connection conn = null;
   counterForInvokingServiceA++;
   conn = connectionService();
   return conn;
}
```

The above code-snippet means the method serviceA applies the `CircuitBreaker` policy, which is to open the circuit once 3 (4x0.75) failures occur among the rolling window of 4 consecutive invocation. The circuit will stay open for 1000ms and then back to half open. After 10 consecutive successful invocations, the circuit will be back to close again.

When a circuit is open, A `CircuitBreakerOpenException` must be thrown. The `@CircuitBreaker` annotation can be used together with `@Timeout`, `@Fallback`, `@Asynchronous`, `@Bulkhead` and `@Retry`. A `@Fallback` can be specified and it will be invoked if the `CircuitBreakerOpenException` is thrown. `@Timeout` can be configured to fail an operation if it takes longer than the `Timeout` value to complete.

Chapter 9. Bulkhead

The `Bulkhead` pattern is to prevent faults in one part of the system from cascading to the entire system, which might bring down the whole system. The implementation is to limit the number of concurrent requests accessing to an instance. Therefore, `Bulkhead` pattern is only effective when applying `@Bulkhead` to a component that can be accessed from multiple contexts.

9.1. Bulkhead Usage

A method or class can be annotated with `@Bulkhead`, which means the method or the methods under the class will have Bulkhead policy applied correspondingly. There are two different approaches to the bulkhead: thread pool isolation and semaphore isolation. When `@Bulkhead` is used with `@Asynchronous`, the thread pool isolation approach will be used. If `@Bulkhead` is used without `@Asynchronous`, the semaphore isolation approach will be used. The thread pool approach allows to configure the maximum concurrent requests together with the waiting queue size. The semaphore approach only allows the concurrent number of requests configuration.

9.1.1. Semaphore style Bulkhead

The below code-snippet means the method serviceA applies the `Bulkhead` policy, which is semaphore approach, limiting the maximum concurrent requests to 5.

```
@Bulkhead(5) // maximum 5 concurrent requests allowed
public Connection serviceA() {
   Connection conn = null;
   counterForInvokingServiceA++;
   conn = connectionService();
   return conn;
}
```

When using the semaphore approach, on reaching maximum request counter, the extra request will fail with `BulkheadException`.

9.1.2. Thread pool style Bulkhead

The below code-snippet means the method serviceA applies the `Bulkhead` policy, which is thread pool approach, limiting the maximum concurrent requests to 5 and the waiting queue size to 8.

```
// maximum 5 concurrent requests allowed, maximum 8 requests allowed in the waiting
queue
@Asynchronous
@Bulkhead(value = 5, waitingTaskQueue = 8)
public Future<Connection> serviceA() {
   Connection conn = null;
   counterForInvokingServiceA++;
   conn = connectionService();
   return CompletableFuture.completedFuture(conn);
}
```

When using the thread pool approach, when a request cannot be added to the waiting queue, `BulkheadException` will be thrown.

The `@Bulkhead` annotation can be used together with `@Fallback`, `@CircuitBreaker`, `@Asynchronous`, `@Timeout` and `@Retry`. If a `@Fallback` is specified, it will be invoked if the `BulkheadException` is thrown.

Chapter 10. Fault Tolerance configuration

This specification defines the programming model to build a resilient microservice. Microservices developed using this feature are guaranteed to be resilient despite of running environments. In some service mash platform, e.g. Istio, has its own Fault Tolerance policy. The operation team might want to use the platform Fault Tolerance. In order to fulfil the requirement, MicroProfile Fault Tolerance provides a capability to have its resilient functionalities except `fallback` disabled. The reason `fallback` is special as the `fallback` business logic can only be defined by microservices not by any other platforms.

Set the config property of `MP_Fault_Tolerance_NonFallback_Enabled` with the value of `false` means the Fault Tolerance is disabled, except `@Fallback`. If the property is absent or with the value of `true`, it means that MicroProfile Fault Tolerance is enabled if any annotations are specified. For more information about how to set config properties, refer to MicroProfile Config specification.

In order to prevent from any unexpected behaviours, the property `MP_Fault_Tolerance_NonFallback_Enabled` will only be read on application starting. Any dynamic changes afterwards will be ignored until the application restarting.

10.1. Config Fault Tolerance parameters

This specification defines the annotations: `@Asynchronous`, `@Bulkhead`, `@CircuitBreaker`, `@Fallback`, `@Retry` and `@Timeout`. Each annotation except `@Asynchronous` has parameters. All of the parameters are configurable. The value of each parameter can be overridden individually or globally.

- Override individual parameters The annotation parameters can be overwritten via config properties in the naming convention of `<classname>/<methodname>/<annotation>/<parameter>`.

In the following code snippet, in order to override the `maxRetries` for serviceB invocation to 100, set the config property `com.acme.test.MyClient/serviceB/Retry/maxRetries=100` Similarly to override the `maxDuration` for ServiceA, set the config property

```
com.acme.test.MyClient/serviceA/Retry/maxDuration=3000
```

- Override parameters globally

If the parameters for a particular annotation need to be configured with the same value for a particular class, use the config property `<classname>/<annotation>/<parameter>` for configuration. For an instance, use the following config property to override all `maxRetries` for `Retry` specified on the class `MyClient` to 100.

```
com.acme.test.MyClient/Retry/maxRetries=100
```

Sometimes, the parameters need to be configured with the same value for the whole microservice. For an instance, all `Timeout` needs to be set to `100ms`. It can be cumbersome to override each occurrence of `Timeout`. In this circumstance, the config property `<annotation>/<parameter>` overrides the corresponding parameter value for the specified annotation. For instance, in order to override the `maxRetries` for the `Retry` to be `30`, specify the config property `Retry/maxRetries=30`.

When multiple config properties are present, the property

`<classname>/<methodname>/<annotation>/<parameter>` takes precedence over `<classname>/<annotation>/<parameter>`, which is followed by `<annotation>/<parameter>`.

The override just changes the value of the corresponding parameter specified in the microservice and nothing more. If no annotation matches the specified parameter, the property will be ignored. For instance, if the annotation `Retry` is specified on the class level for the class `com.acme.ClassA`, which has method methodB, the config property `com.acme.ClassA/methodB/Retry/maxRetries` will be ignored. In order to override the property, the config property `com.acme.ClassA/Retry/maxRetries` or `Retry/maxRetries` needs to be specified.

```
package come.acme.test;
public class MyClient{
    /**
     * The configured the max retries is 90 but the max duration is 1000ms.
     * Once the duration is reached, no more retries should be performed,
     * even through it has not reached the max retries.
     */
    @Retry(maxRetries = 90, maxDuration= 1000)
    public void serviceB() {
        writingService();
    }

    /**
    * There should be 0-800ms (jitter is -400ms - 400ms) delays
    * between each invocation.
    * there should be at least 4 retries but no more than 10 retries.
    */
    @Retry(delay = 400, maxDuration= 3200, jitter= 400, maxRetries = 10)
    public Connection serviceA() {
        return connectionService();
    }

    /**
    * Sets retry condition, which means Retry will be performed on
    * IOException.
    */
    @Retry(retryOn = {IOException.class})
    public void serviceB() {
        writingService();
    }

}
```

If an annotation is not present, the configured properties are ignored. For instance, the property `com.acme.ClassA/methodB/Retry/maxRetries` will be ignored if `@Retry` annotation is not specified on the `methodB` of `com.acme.ClassA`. Similarly, the property `com.acme.ClassA/Retry/maxRetries` will be ignored if `@Retry` annotation is not specified on the class `com.acme.ClassA` as a class-level annotation.