主讲老师：Fox

1. 前言

课前思考：

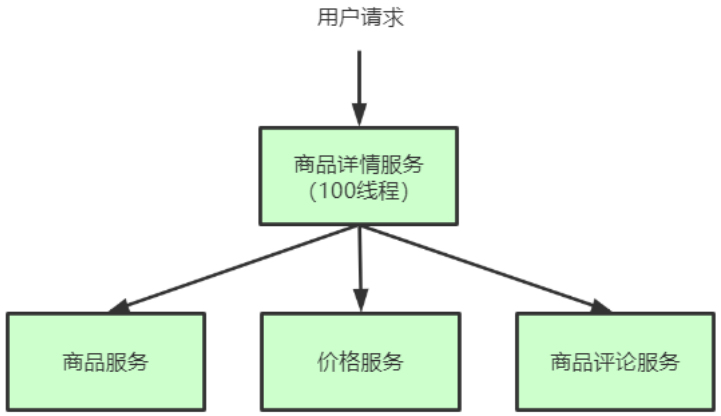
1、当服务访问量达到一定程度，流量扛不住的时候，该如何处理？

2、服务之间相互依赖，当服务A出现响应时间过长，影响到服务B的响应，进而产生连锁反应，直至影响整个依赖链上的所有服务，该如何处理？

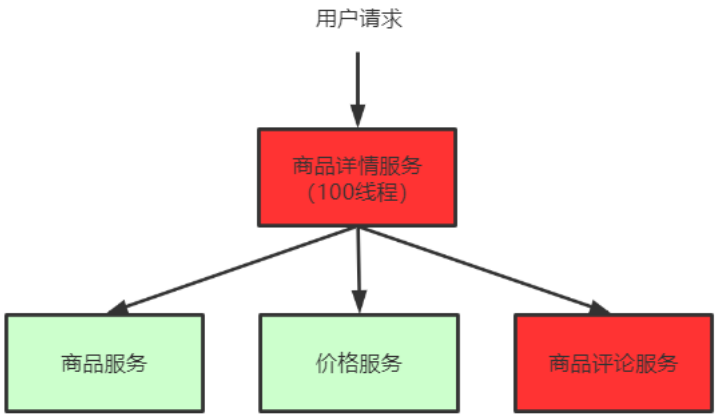
这是分布式、微服务开发不可避免的问题。

1.1 分布式系统遇到的问题

在一个高度服务化的系统中,我们实现的一个业务逻辑通常会依赖多个服务,比如:商品详情展示服务会依赖商品服务, 价格服务, 商品评论服务. 如图所示:



调用三个依赖服务会共享商品详情服务的线程池. 如果其中的商品评论服务不可用, 就会出现线程池里所有线程都因等待响应而被阻塞, 从而造成服务雪崩. 如图所示:



服务雪崩效应：因服务提供者的不可用导致服务调用者的不可用,并将不可用逐渐放大的过程，就叫服务雪崩效应

导致服务不可用的原因： 程序Bug，大流量请求，硬件故障，缓存击穿

【大流量请求】：在秒杀和大促开始前,如果准备不充分,瞬间大量请求会造成服务提供者的不可用。

【硬件故障】：可能为硬件损坏造成的服务器主机宕机, 网络硬件故障造成的服务提供者的不可访问。

【缓存击穿】：一般发生在缓存应用重启, 缓存失效时高并发， 所有缓存被清空时,以及短时间内大量缓存失效时。大量的缓存不命中, 使请求直击后端,造成服务提供者超负荷运行,引起服务不可用。

在服务提供者不可用的时候，会出现大量重试的情况：用户重试、代码逻辑重试，这些重试最终导致：进一步加大请求流量。所以归根结底导致雪崩效应的最根本原因是：大量请求线程同步等待造成的资源耗尽。当服务调用者使用同步调用时, 会产生大量的等待线程占用系统资源。一旦线程资源被耗尽,服务调用者提供的服务也将处于不可用状态, 于是服务雪崩效应产生了。

1.2 解决方案

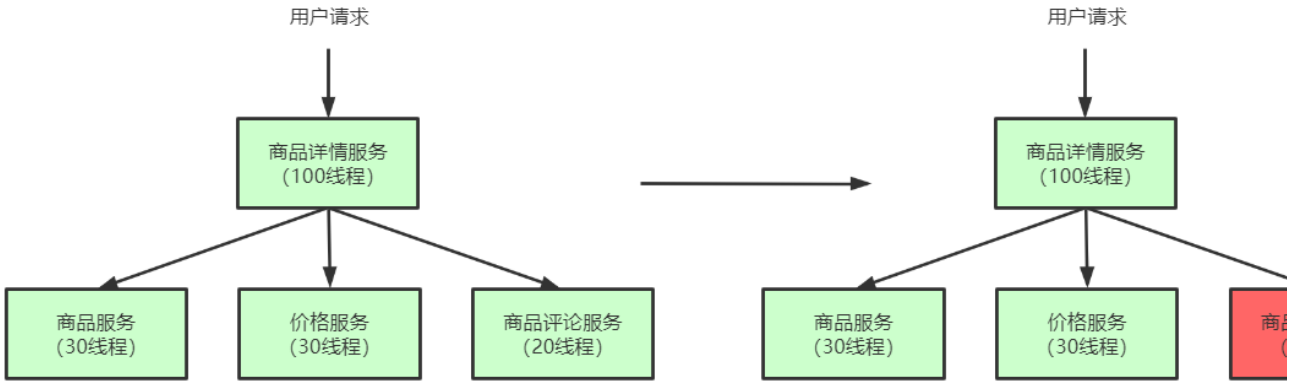
超时机制

在不做任何处理的情况下，服务提供者不可用会导致消费者请求线程强制等待，而造成系统资源耗尽。加入超时机制，一旦超时，就释放资源。由于释放资源速度较快，一定程度上可以抑制资源耗尽的问题。

服务限流(资源隔离)

限制请求核心服务提供者的流量，使大流量拦截在核心服务之外，这样可以更好的保证核心服务提供者不出问题，对于一些出问题的服务可以限制流量访问，只分配固定线程资源访问，这样能使整体的资源不至于被出问题的服务耗尽，进而整个系统雪崩。那么服务之间怎么限流，怎么资源隔离？例如可以通过线程池+队列的方式，通过信号量的方式。

如下图所示, 当商品评论服务不可用时, 即使商品服务独立分配的20个线程全部处于同步等待状态,也不会影响其他依赖服务的调用。



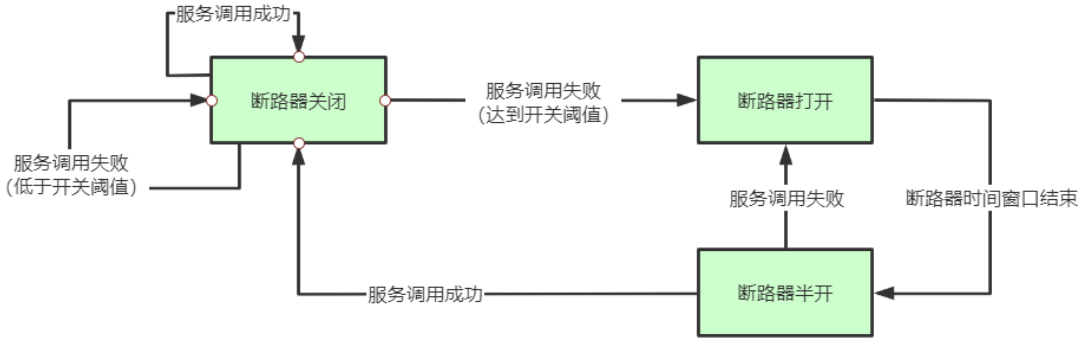
服务熔断

远程服务不稳定或网络抖动时暂时关闭，就叫服务熔断。

现实世界的断路器大家肯定都很了解，断路器实时监控电路的情况，如果发现电路电流异常，就会跳闸，从而防止电路被烧毁。

软件世界的断路器可以这样理解：实时监测应用，如果发现在一定时间内失败次数/失败率达到一定阈值，就“跳闸”，断路器打开——此时，请求直接返回，而不去调用原本调用的逻辑。跳闸一段时间后（例如10秒），断路器会进入半开状态，这是一个瞬间态，此时允许一次请求调用该调的逻辑，如果成功，则断路器关闭，应用正常调用；如果调用依然不成功，断路器继续回到打开状态，过段时间再进入半开状态尝试——通过”跳闸“，应用可以保护自己，而且避免浪费资源；而通过半开的设计，可实现应用的“自我修复“。

所以，同样的道理，当依赖的服务有大量超时时，在让新的请求去访问根本没有意义，只会无畏的消耗现有资源。比如我们设置了超时时间为1s,如果短时间内有大量请求在1s内都得不到响应，就意味着这个服务出现了异常，此时就没有必要再让其他的请求去访问这个依赖了，这个时候就应该使用断路器避免资源浪费。



服务降级

有服务熔断，必然要有服务降级。

所谓降级，就是当某个服务熔断之后，服务将不再被调用，此时客户端可以自己准备一个本地的fallback（回退）回调，返回一个缺省值。 例如：(备用接口/缓存/mock数据) 。这样做，虽然服务水平下降，但好歹可用，比直接挂掉要强，当

然这也要看适合的业务场景。

2. Sentinel: 分布式系统的流量防卫兵

2.1 Sentinel 是什么

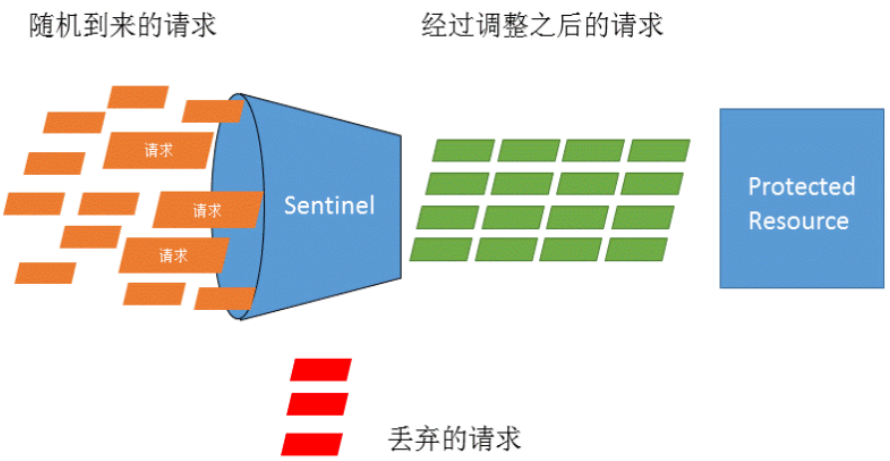
随着微服务的流行，服务和服务之间的稳定性变得越来越重要。Sentinel 是面向分布式服务架构的流量控制组件，主要以流量为切入点，从限流、流量整形、熔断降级、系统负载保护、热点防护等多个维度来帮助开发者保障微服务的稳定性。

源码地址：https://github.com/alibaba/Sentinel

官方文档：https://github.com/alibaba/Sentinel/wiki

Sentinel具有以下特征:

- **丰富的应用场景**： Sentinel 承接了阿里巴巴近 10 年的双十一大促流量的核心场景，例如秒杀（即突发流量控制在系统容量可以承受的范围）、消息削峰填谷、实时熔断下游不可用应用等。
- **完备的实时监控**： Sentinel 同时提供实时的监控功能。您可以在控制台中看到接入应用的单台机器秒级数据，甚至 500 台以下规模的集群的汇总运行情况。
- **广泛的开源生态**： Sentinel 提供开箱即用的与其它开源框架/库的整合模块，例如与 Spring Cloud、Dubbo、gRPC 的整合。您只需要引入相应的依赖并进行简单的配置即可快速地接入 Sentinel。
- **完善的 SPI 扩展点**： Sentinel 提供简单易用、完善的 SPI 扩展点。您可以通过实现扩展点，快速的定制逻辑。例如定制规则管理、适配数据源等。



Sentinel和Hystrix对比

https://github.com/alibaba/Sentinel/wiki/Sentinel-%E4%B8%8E-Hystrix-%E7%9A%84%E5%AF%B9%E6%AF%94

	Sentinel	Hystrix
隔离策略	信号量隔离	线程池隔离/信号量隔离
熔断降级策略	基于响应时间或失败比率	基于失败比率
实时指标实现	滑动窗口	滑动窗口（基于 RxJava）
规则配置	支持多种数据源	支持多种数据源
扩展性	多个扩展点	插件的形式
基于注解的支持	支持	支持
限流	基于 QPS，支持基于调用关系的限流	有限的支持
流量整形	支持慢启动、匀速器模式	不支持
系统负载保护	支持	不支持
控制台	开箱即用，可配置规则、查看秒级监控、机器发现等	不完善
常见框架的适配	Servlet、Spring Cloud、Dubbo、gRPC 等	Servlet、Spring Cloud Netflix

2.2 Sentinel 工作原理

2.2.1 基本概念

资源

资源是 Sentinel 的关键概念。它可以是 Java 应用程序中的任何内容，例如，由应用程序提供的服务，或由应用程序调用的其它应用提供的服务，甚至可以是一段代码。在接下来的文档中，我们都会用资源来描述代码块。

只要通过 Sentinel API 定义的代码，就是资源，能够被 Sentinel 保护起来。大部分情况下，可以使用方法签名，URL，甚至服务名称作为资源名来标示资源。

规则

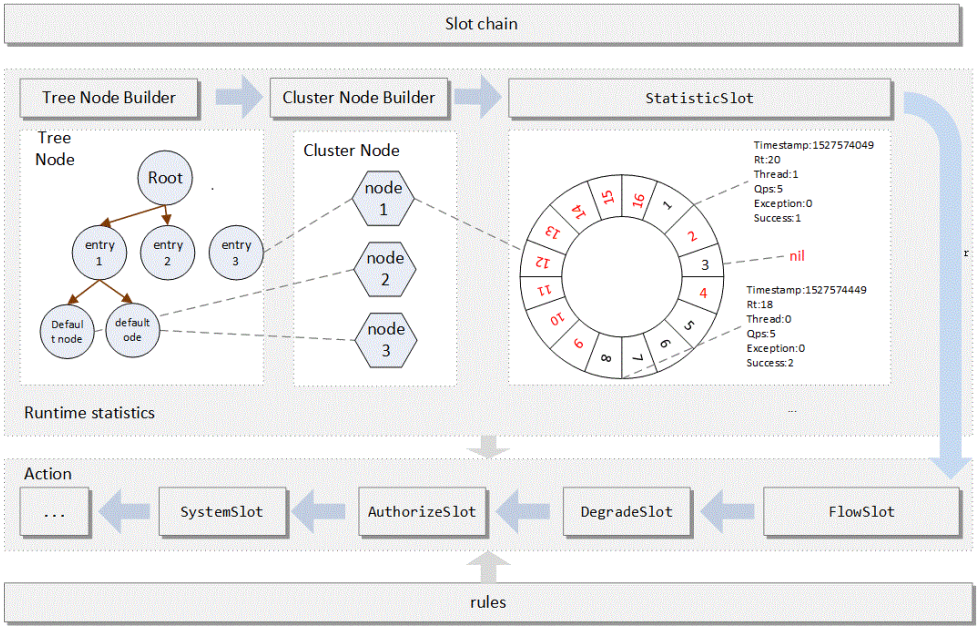
围绕资源的实时状态设定的规则，可以包括流量控制规则、熔断降级规则以及系统保护规则。所有规则可以动态实时调整。

2.2.2 Sentinel工作主流程

https://github.com/alibaba/Sentinel/wiki/Sentinel%E5%B7%A5%E4%BD%9C%E4%B8%BB%E6%B5%81%E7%A8%8B

在 Sentinel 里面，所有的资源都对应一个资源名称（`resourceName`），每次资源调用都会创建一个 `Entry` 对象。Entry 可以通过对主流框架的适配自动创建，也可以通过注解的方式或调用 `SphU` API 显式创建。Entry 创建的时候，同时也会创建一系列功能插槽（slot chain），这些插槽有不同的职责，例如:

- `NodeSelectorSlot` 负责收集资源的路径，并将这些资源的调用路径，以树状结构存储起来，用于根据调用路径来限流降级；
- `ClusterBuilderSlot` 则用于存储资源的统计信息以及调用者信息，例如该资源的 RT, QPS, thread count 等等，这些信息将用作为多维度限流，降级的依据；
- `StatisticSlot` 则用于记录、统计不同纬度的 runtime 指标监控信息；
- `FlowSlot` 则用于根据预设的限流规则以及前面 slot 统计的状态，来进行流量控制；
- `AuthoritySlot` 则根据配置的黑白名单和调用来源信息，来做黑白名单控制；
- `DegradeSlot` 则通过统计信息以及预设的规则，来做熔断降级；
- `SystemSlot` 则通过系统的状态，例如 load1 等，来控制总的入口流量；



2.3 Sentinel快速开始

在官方文档中，定义的Sentinel进行资源保护的几个步骤：

1. 定义资源
2. 定义规则
3. 检验规则是否生效

```
Entry entry = null;
// 务必保证 finally 会被执行
try {
 // 资源名可使用任意有业务语义的字符串
 entry = SphU.entry("自定义资源名");
 // 被保护的业务逻辑
 // do something...
} catch (BlockException ex) {
 // 资源访问阻止，被限流或被降级
 // 进行相应的处理操作
 } catch (Exception ex) {
 // 若需要配置降级规则，需要通过这种方式记录业务异常
 Tracer.traceEntry(ex, entry);
} finally {
 // 务必保证 exit，务必保证每个 entry 与 exit 配对
 if (entry != null) {
 entry.exit();
 }
```

Sentinel资源保护的方式

API实现

1. 引入依赖

```
<dependency>
 <groupId>com.alibaba.csp</groupId>
 <artifactId>sentinel-core</artifactId>
 <version>1.8.0</version>
</dependency>
```

2. 编写测试逻辑

```
@RestController
@Slf4j
public class HelloController {

 private static final String RESOURCE_NAME = "hello";
```

```java

@RequestMapping(value = "/hello")
public String hello() {

Entry entry = null;
try {
// 资源名可使用任意有业务语义的字符串，比如方法名、接口名或其它可唯一标识的字符串。
entry = SphU.entry(RESOURCE_NAME);
// 被保护的业务逻辑
String str = "hello world";
log.info("====="+str);
return str;
} catch (BlockException e1) {
// 资源访问阻止，被限流或被降级
//进行相应的处理操作
log.info("block!");
} catch (Exception ex) {
// 若需要配置降级规则，需要通过这种方式记录业务异常
Tracer.traceEntry(ex, entry);
} finally {
if (entry != null) {
entry.exit();
}
}
return null;
}

/**
* 定义流控规则
*/
@PostConstruct
private static void initFlowRules(){
List<FlowRule> rules = new ArrayList<>();
FlowRule rule = new FlowRule();
//设置受保护的资源
rule.setResource(RESOURCE_NAME);
// 设置流控规则 QPS
rule.setGrade(RuleConstant.FLOW_GRADE_QPS);
// 设置受保护的资源阈值
// Set limit QPS to 20.
rule.setCount(1);
rules.add(rule);
// 加载配置好的规则
FlowRuleManager.loadRules(rules);
}
}
```

测试效果：

```
er.HelloController        : =====hello world
er.HelloController        : =====hello world
er.HelloController        : block!
er.HelloController        : =====hello world
er.HelloController        : block!
er.HelloController        : block!
er.HelloController        : block!
er.HelloController        : block!
er.HelloController        : =====hello world
```

缺点：

- 业务侵入性很强，需要在controller中写入非业务代码.
- 配置不灵活 若需要添加新的受保护资源 需要手动添加 init方法来添加流控规则

@SentinelResource注解实现

> `@SentinelResource` 注解用来标识资源是否被限流、降级。
> blockHandler：定义当资源内部发生了BlockException应该进入的方法（捕获的是Sentinel定义的异常）
> fallback：定义的是资源内部发生了Throwable应该进入的方法
> exceptionsToIgnore：配置fallback可以忽略的异常
> 源码入口：com.alibaba.csp.sentinel.annotation.aspectj.SentinelResourceAspect

1.引入依赖

```
<dependency>
 <groupId>com.alibaba.csp</groupId>
 <artifactId>sentinel-annotation-aspectj</artifactId>
 <version>1.8.0</version>
</dependency>
```

2.配置切面支持

```
@Configuration
public class SentinelAspectConfiguration {

 @Bean
 public SentinelResourceAspect sentinelResourceAspect() {
 return new SentinelResourceAspect();
 }
}
```

3.UserController中编写测试逻辑，添加@SentinelResource，并配置blockHandler和fallback

```
@RequestMapping(value = "/findOrderByUserId/{id}")
@SentinelResource(value = "findOrderByUserId",
 fallback = "fallback",fallbackClass = ExceptionUtil.class,
 blockHandler = "handleException",blockHandlerClass = ExceptionUtil.class
 )
public R findOrderByUserId(@PathVariable("id") Integer id) {
 //ribbon实现
 String url = "http://mall-order/order/findOrderByUserId/"+id;
 R result = restTemplate.getForObject(url,R.class);

 if(id==4){
 throw new IllegalArgumentException("非法参数异常");
 }

 return result;
}
```

4.编写ExceptionUtil，注意如果指定了class，方法必须是static方法

```
public class ExceptionUtil {

 public static R fallback(Integer id,Throwable e){
 return R.error(-2,"===被异常降级啦===");
 }

 public static R handleException(Integer id, BlockException e){
 return R.error(-2,"===被限流啦===");
 }
}
```

5.流控规则设置可以通过Sentinel dashboard配置

客户端需要引入 Transport 模块来与 Sentinel 控制台进行通信。

```
<dependency>
 <groupId>com.alibaba.csp</groupId>
 <artifactId>sentinel-transport-simple-http</artifactId>
 <version>1.8.0</version>
</dependency>
```

6. 启动 Sentinel 控制台

下载控制台 jar 包并在本地启动：可以参见 此处文档

```
#启动控制台命令
java -jar sentinel-dashboard-1.8.0.jar
```

用户可以通过如下参数进行配置：

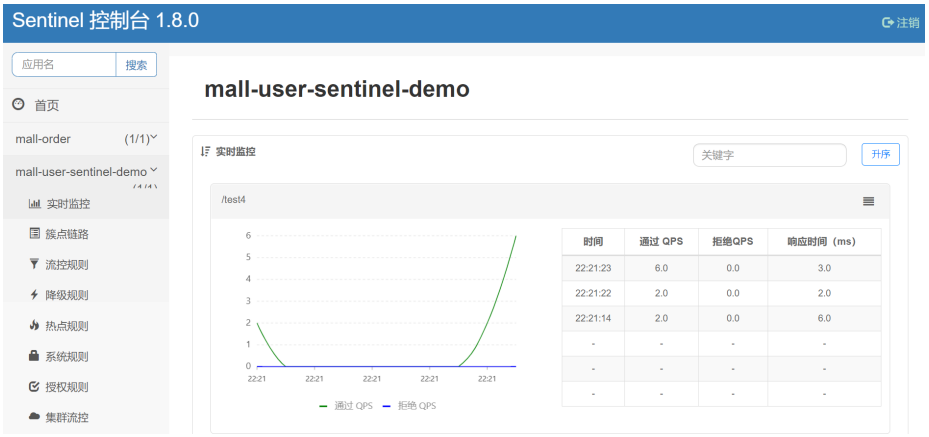
`-Dsentinel.dashboard.auth.username=sentinel` 用于指定控制台的登录用户名为 `sentinel`；

`-Dsentinel.dashboard.auth.password=123456` 用于指定控制台的登录密码为 `123456`；如果省略这两个参数，默认用户和密码均为 `sentinel`；

`-Dserver.servlet.session.timeout=7200` 用于指定 Spring Boot 服务端 session 的过期时间，如 `7200` 表示 7200 秒；`60m` 表示 60 分钟，默认为 30 分钟；

访问http://localhost:8080/#/login ,默认用户名密码： sentinel/sentinel

Sentinel 会在客户端首次调用的时候进行初始化，开始向控制台发送心跳包，所以要确保客户端有访问量；



2.4 Spring Cloud Alibaba整合Sentinel

1.引入依赖

```
<dependency>
  <groupId>com.alibaba.cloud</groupId>
  <artifactId>spring-cloud-starter-alibaba-sentinel</artifactId>
</dependency>

<dependency>
  <groupId>org.springframework.boot</groupId>
  <artifactId>spring-boot-starter-actuator</artifactId>
</dependency>
```

2.添加yml配置，为微服务设置sentinel控制台地址

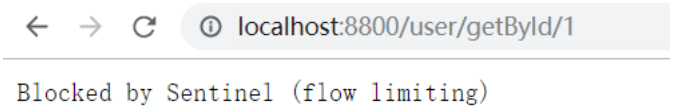
添加Sentinel后，需要暴露/actuator/sentinel端点,而Springboot默认是没有暴露该端点的，所以需要设置，测试http://localhost:8800/actuator/sentinel

```yaml
server:
 port: 8800

spring:
 application:
 name: mall-user-sentinel-demo
 cloud:
 nacos:
 discovery:
 server-addr: 127.0.0.1:8848

 sentinel:
 transport:
 # 添加sentinel的控制台地址
 dashboard: 127.0.0.1:8080
 # 指定应用与Sentinel控制台交互的端口，应用本地会起一个该端口占用的HttpServer
 # port: 8719

#暴露actuator端点
management:
 endpoints:
 web:
 exposure:
 include: '*'
```

3.在sentinel控制台中设置流控规则

- **资源名**：接口的API
- **针对来源**：默认是default，当多个微服务都调用这个资源时，可以配置微服务名来对指定的微服务设置阈值
- **阈值类型**：分为QPS和线程数 假设阈值为10
- **QPS类型**：只得是每秒访问接口的次数>10就进行限流
- **线程数**：为接受请求该资源分配的线程数>10就进行限流

测试： 因为QPS是1，所以1秒内多次访问会出现如下情形：



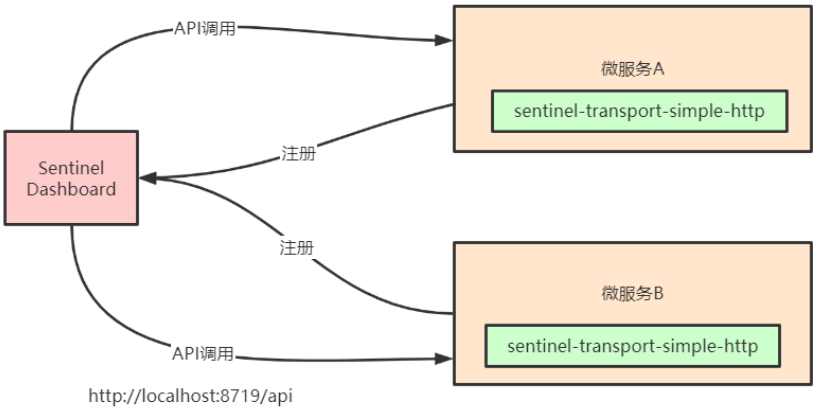

```
Blocked by Sentinel (flow limiting)
```

访问http://localhost:8800/actuator/sentinel， 可以查看flowRules

```
{
    blockPage: null,
    appName: "service-sentinel-consumer",
    consoleServer: "localhost:8888",
    coldFactor: "3",
  - rules: {
        systemRules: [ ],
        authorityRule: [ ],
        paramFlowRule: [ ],
      - flowRules: [
          - {
                resource: "/user/getById/1",
                limitApp: "default",
                grade: 1,
                count: 1,
                strategy: 0,
                refResource: null,
                controlBehavior: 0,
                warmUpPeriodSec: 10,
                maxQueueingTimeMs: 500,
                clusterMode: false,
              - clusterConfig: {
                    flowId: null,
                    thresholdType: 0,
                    fallbackToLocalWhenFail: true,
                    strategy: 0,
                    sampleCount: 10,
                    windowIntervalMs: 1000
                }
            }
        ],
        degradeRules: [ ]
    }
```


微服务和Sentinel Dashboard通信原理

Sentinel控制台与微服务端之间，实现了一套服务发现机制，集成了Sentinel的微服务都会将元数据传递给Sentinel控制台，架构图如下所示：



文档：08 微服务组件Sentinel实战.note

链接：http://note.youdao.com/noteshare?id=2dcecdcc67311fe752754b252bd457c2&sub=8087295E17A34E07A7C77111D0A2DADF