20 | 服务端出现故障时该如何应对？

2018-10-06 胡忠想

从0开始学微服务 进入课程 >

讲述：胡忠想
时长 09:49 大小 4.51M

在专栏前面我讲过，单体应用改造成微服务的一个好处是可以减少故障影响范围，故障被局限在一个微服务系统本身，而不是整个单体应用都崩溃。那么具体到一个微服务系统，如果出现了故障，应该如何处理呢？

首先，我先来带你回顾一下微服务系统可能出现故障的种类，主要有三种故障。

集群故障。根据我的经验，微服务系统一般都是集群部署的，根据业务量大小而定，集群规模从几台到甚至上万台都有可能。一旦某些代码出现 bug，可能整个集群都会发生故障，不能提供对外提供服务。

单 IDC 故障。现在大多数互联网公司为了保证业务的高可用性，往往业务部署在不止一个 IDC。然而现实中时常会发生某个 IDC 的光缆因为道路施工被挖断，导致整个 IDC 脱网。

单机故障。顾名思义就是集群中的个别机器出现故障，这种情况往往对全局没有太大影响，但会导致调用到故障机器上的请求都失败，影响整个系统的成功率。

在我的实践过程中，这三种故障都经常遇到，因此相应的处理手段也可谓驾轻就熟，下面就把我应对故障的实战经验分享给你，希望对你有所帮助。

集群故障

一般而言，集群故障的产生原因不外乎有两种：一种是代码 bug 所导致，比如说某一段 Java 代码不断地分配大对象，但没有及时回收导致 JVM OOM 退出；另一种是突发的流量冲击，超出了系统的最大承载能力，比如“双 11”这种购物活动，电商系统会在零点一瞬间涌入大量流量，超出系统的最大承载能力，一下子就把整个系统给压垮了。

应付集群故障的思路，主要有两种：**限流**和**降级**。

1. 限流

顾名思义，限流就是限制流量，通常情况下，系统能够承载的流量根据集群规模的大小是固定的，可以称之为系统的最大容量。当真实流量超过了系统的最大容量后，就会导致系统响应变慢，服务调用出现大量超时，反映给用户的感觉就是卡顿、无响应。所以，应该根据系统的最大容量，给系统设置一个阈值，超过这个阈值的请求会被自动抛弃，这样的话可以最大限度地保证系统提供的服务正常。

除此之外，通常一个微服务系统会同时提供多个服务，每个服务在同一时刻的请求量也是不同的，很可能出现的一种情况就是，系统中某个服务的请求量突增，占用了系统中大部分资源，导致其他服务没有资源可用。因此，还要针对系统中每个服务的请求量也设置一个阈值，超过这个阈值的请求也要被自动抛弃，这样的话不至于因为一个服务影响了其他所有服务。

在实际项目中，可以用两个指标来衡量服务的请求量，一个是 QPS 即每秒请求量，一个是工作线程数。不过 QPS 因为不同服务的响应快慢不同，所以系统能够承载的 QPS 相差很大，因此一般选择工作线程数来作为限流的指标，给系统设置一个总的最大工作线程数以及单个服务的最大工作线程数，这样的话无论是系统的总请求量过大导致整体工作线程数量达到最大工作线程数，还是某个服务的请求量超过单个服务的最大工作线程数，都会被限流，以起到保护整个系统的作用。

2. 降级

什么是降级呢？在我看来，降级就是通过停止系统中的某些功能，来保证系统整体的可用性。降级可以说是一种被动防御的措施，为什么这么说呢？因为它一般是系统已经出现故障后所采取的一种止损措施。

那么降级一般是如何实现的呢？根据我的实践来看， 一种可行的方案是通过开关来实现。

具体来讲，就是在系统运行的内存中开辟一块区域，专门用于存储开关的状态，也就是开启还是关闭。并且需要监听某个端口，通过这个端口可以向系统下发命令，来改变内存中开关的状态。当开关开启时，业务的某一段逻辑就不再执行，而正常情况下，开关是关闭的状态。

开关一般用在两种地方，一种是新增的业务逻辑，因为新增的业务逻辑相对来说不成熟，往往具备一定的风险，所以需要加开关来控制新业务逻辑是否执行；另一种是依赖的服务或资源，因为依赖的服务或者资源不总是可靠的，所以最好是有开关能够控制是否对依赖服务或资源发起调用，来保证即使依赖出现问题，也能通过降级来避免影响。

在实际业务应用的时候，降级要按照对业务的影响程度进行分级，一般分为三级：一级降级是对业务影响最小的降级，在故障的情况下，首先执行一级降级，所以一级降级也可以设置成自动降级，不需要人为干预；二级降级是对业务有一定影响的降级，在故障的情况下，如果一级降级起不到多大作用的时候，可以人为采取措施，执行二级降级；三级降级是对业务有较大影响的降级，这种降级要么是对商业收入有重大影响，要么是对用户体验有重大影响，所以操作起来要非常谨慎，不在最后时刻一般不予采用。

单 IDC 故障

在现实情况下，整个 IDC 脱网的事情时有发生，多半是因为不可抗力比如机房着火、光缆被挖断等，如果业务全部部署在这个 IDC，那就完全不可访问了，所以国内大部分的互联网业务多采用多 IDC 部署。具体来说，有的采用同城双活，也就是在一个城市的两个 IDC 内部署；有的采用异地多活，一般是在两个城市的两个 IDC 内部署；当然也有支付宝这种金融级别的应用采用了“三地五中心”部署，这种部署成本显然高比两个 IDC 要高得多，但可用性的保障要更高。

采用多 IDC 部署的最大好处就是当有一个 IDC 发生故障时，可以把原来访问故障 IDC 的流量切换到正常的 IDC，来保证业务的正常访问。

流量切换的方式一般有两种，一种是基于 DNS 解析的流量切换，一种是基于 RPC 分组的流量切换。

1. 基于 DNS 解析的流量切换

基于 DNS 解析流量的切换，一般是通过把请求访问域名解析的 VIP 从一个 IDC 切换到另外一个 IDC。比如访问“www.weibo.com”，正常情况下北方用户会解析到联通机房的 VIP，南方用户会解析到电信机房的 VIP，如果联通机房发生故障的话，会把北方用户访问也解析到电信机房的 VIP，只不过此时网络延迟可能会变长。

2. 基于 RPC 分组的流量切换

对于一个服务来说，如果是部署在多个 IDC 的话，一般每个 IDC 就是一个分组。假如一个 IDC 出现故障，那么原先路由到这个分组的流量，就可以通过向配置中心下发命令，把原先路由到这个分组的流量全部切换到别的分组，这样的话就可以切换故障 IDC 的流量了。

单机故障

单机故障是发生概率最高的一种故障了，尤其对于业务量大的互联网应用来说，上万台机器的规模也是很常见的。这种情况下，发生单机故障的概率就很高了，这个时候只靠运维人肉处理显然不可行，所以就要求有某种手段来自动处理单机故障。

根据我的经验，处理单机故障一个有效的办法就是自动重启。具体来讲，你可以设置一个阈值，比如以某个接口的平均耗时为准，当监控单机上某个接口的平均耗时超过一定阈值时，就认为这台机器有问题，这个时候就需要把有问题的机器从线上集群中摘除掉，然后在重启服务后，重新加入到集群中。

不过这里要注意的是，需要防止网络抖动造成的接口超时从而触发自动重启。一种方法是在收集单机接口耗时数据时，多采集几个点，比如每 10s 采集一个点，采集 5 个点，当 5 个点中有超过 3 个点的数据都超过设定的阈值范围，才认为是真正的单机问题，这时会触发自动重启策略。

除此之外，为了防止某些特殊情况下，短时间内被重启的单机过多，造成整个服务池可用节点数太少，最好是设置一个可重启的单机数量占整个集群的最大比例，一般这个比例不要超过 10%，因为正常情况下，不大可能有超过 10% 的单机都出现故障。

总结

今天我们探讨了微服务系统可能出现的三种故障：集群故障、单 IDC 故障、单机故障，并且针对这三种故障我给出了分别的解决方案，包括降级、限流、流量切换以及自动重启。

在遇到实际的故障时，往往多个手段是并用的，比如在出现单 IDC 故障，首先要快速切换流量到正常的 IDC，但此时可能正常 IDC 并不足以支撑两个 IDC 的流量，所以这个时候首先要降级部分功能，保证正常的 IDC 顺利支撑切换过来的流量。

而且要尽量让故障处理自动化，这样可以大大减少故障影响的时间。因为一旦需要引入人为干预，往往故障处理的时间都得是 10 分钟以上，这对大部分用户敏感型业务的影响是巨大的，如果能做到自动化故障处理的话，可以将故障处理的时间降低到 1 分钟以内甚至秒级别，这样的话对于用户的影响最小。

思考题

上面我提到为了避免单 IDC 故障导致服务不可用情况的发生，服务需要采用多 IDC 部署，这个时候就要求服务依赖的数据也需要存储在多个 IDC 内，这样势必会带来数据一致性的问题，你有什么解决方案吗？

欢迎你在留言区写下自己的思考，与我一起讨论。

上一篇　19 | 如何使用服务路由？

下一篇　21 | 服务调用失败时有哪些处理手段？

精选留言 (15)

写留言

叽歪
2018-10-10

29

文章深度不够，能不能在深入一点

展开

叽歪
2018-10-10

4

单机故障重启，这个没有说服力，故障是肯定有原因的，应该找到根本原因.故障重启不是一个资深工程师的解决问题的方法.强烈不认同

展开

作者回复: 单机故障有多种原因，一切以恢复线上服务为首要目标，重启后可以发邮件告警，如果有机器出现频繁重启，运维就需要介入查看具体原因

拉欧
2018-10-09

2

只能通过最终一致性保证，比如MySQL的binlog复制

展开

波波安
2018-10-14

1

一般不保证实时一致性，只能保证最终一致性。
1、两个中心中拉专线，进行数据底层同步。2、对于重要数据在做完底层同步后还可以通过消息队列再做一次异步同步来作为补偿。但是要控制好，防止数据重复写入。

展开

公号-代码...

2018-10-06

1

拉光纤专线能否实现？但是数据完全一致似乎很难做到？

好好学习

2018-10-06

1

单机故障，重启过于暴力，硬件问题出现ping不可达，假死可以重启，其他软件层面还是要有其他层面的分析

作者回复: 是的，可以设置最大重启次数，超过这个就不重启了

张小男

2019-04-16

最大线线程数设置多少合理？我们高峰期有10000多，当然服务现在还没拆！这样8核16G都跑不满

张小男

2019-04-16

线程数达到多少算瓶颈？我们高峰期现在是10000多，8核16G！cpu都跑不满

松花皮蛋me

2019-02-17

负载均衡中健康检测可以应对单机故障的情况

展开

花生

2019-01-13

用九桥，在异地机房在数据实时备份

展开

Liam

2018-10-18

能否讲讲具体的限流方法，例如有接入层限流，nginx的配置等

作者回复: 限流有好几种算法，比如令牌痛算法，可以搜搜它的原理

大光头

2018-10-08

一致性分为弱一致性和强一致性。强一致性比如银行转账业务，就必须要求是强一致性。而对于微博评论，并不要求发出去之后立马看到，只要最终看到就可以，晚一点也没关系。
对于强一致性系统，一般都会有读写分离，读可以从多个备份中取数据，而写必须要数据同步到所有备份之后返回。...

展开

malasang

2018-10-08

分布式数据库，通过分布式锁控制数据写入和更新删除。

明天更美好

2018-10-07

异步不复制，最终一致性即可

展开

vick

2018-10-07

如果没有很大实时性需求的话，两个idc之间保持后台数据同步，故障恢复后等数据同步完成再开启关闭的idc。