# 25 | 高可用存储架构：双机架构

2018-06-23 李运华

25 | 高可用存储架构：双机架构
朗读人：黄洲君　15′29″ | 7.10M

存储高可用方案的本质都是通过将数据复制到多个存储设备，通过数据冗余的方式来实现高可用，其复杂性主要体现在如何应对复制延迟和中断导致的数据不一致问题。因此，对任何一个高可用存储方案，我们需要从以下几个方面去进行思考和分析：

- 数据如何复制？
- 各个节点的职责是什么？
- 如何应对复制延迟？
- 如何应对复制中断？

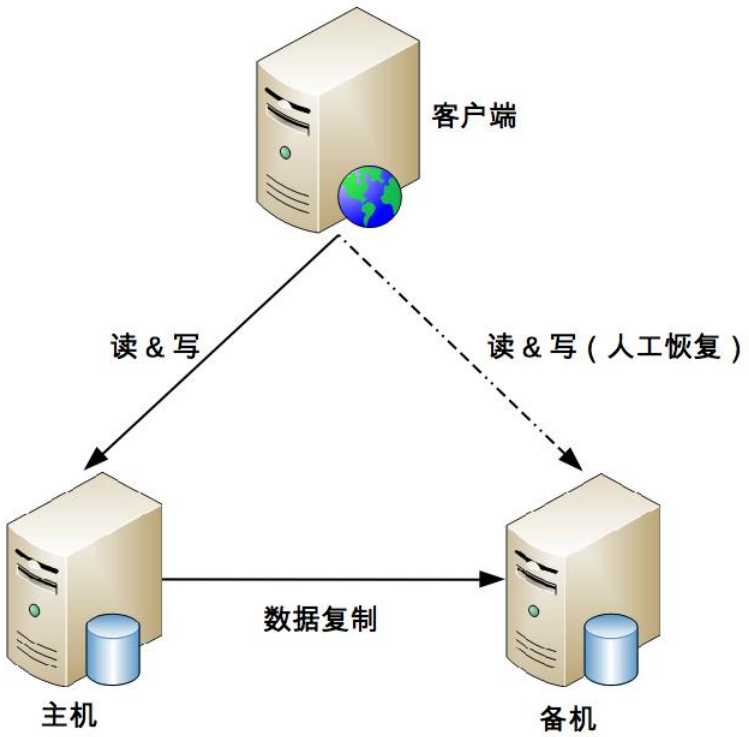
常见的高可用存储架构有主备、主从、主主、集群、分区，每一种又可以根据业务的需求进行一些特殊的定制化功能，由此衍生出更多的变种。由于不同业务的定制功能难以通用化，今天我将针对业界通用的方案，来分析常见的双机高可用架构：主备、主从、主备 / 主从切换和主主。

## 主备复制

主备复制是最常见也是最简单的一种存储高可用方案，几乎所有的存储系统都提供了主备复制的功能，例如 MySQL、Redis、MongoDB 等。

1. 基本实现

下面是标准的主备方案结构图：




其整体架构比较简单，主备架构中的“备机”主要还是起到一个备份作用，并不承担实际的业务读写操作，如果要把备机改为主机，需要人工操作。

2. 优缺点分析

主备复制架构的优点就是简单，表现有：

- 对于客户端来说，不需要感知备机的存在，即使灾难恢复后，原来的备机被人工修改为主机后，对于客户端来说，只是认为主机的地址换了而已，无须知道是原来的备机升级为主机。
- 对于主机和备机来说，双方只需要进行数据复制即可，无须进行状态判断和主备切换这类复杂的操作。

主备复制架构的缺点主要有：

- 备机仅仅只为备份，并没有提供读写操作，硬件成本上有浪费。
- 故障后需要人工干预，无法自动恢复。人工处理的效率是很低的，可能打电话找到能够操作的人就耗费了 10 分钟，甚至如果是深更半夜，出了故障都没人知道。人工在执行恢复操作的过程中也容易出错，因为这类操作并不常见，可能 1 年就 2、3 次，实际操作的时候很可能遇到各种意想不到的问题。

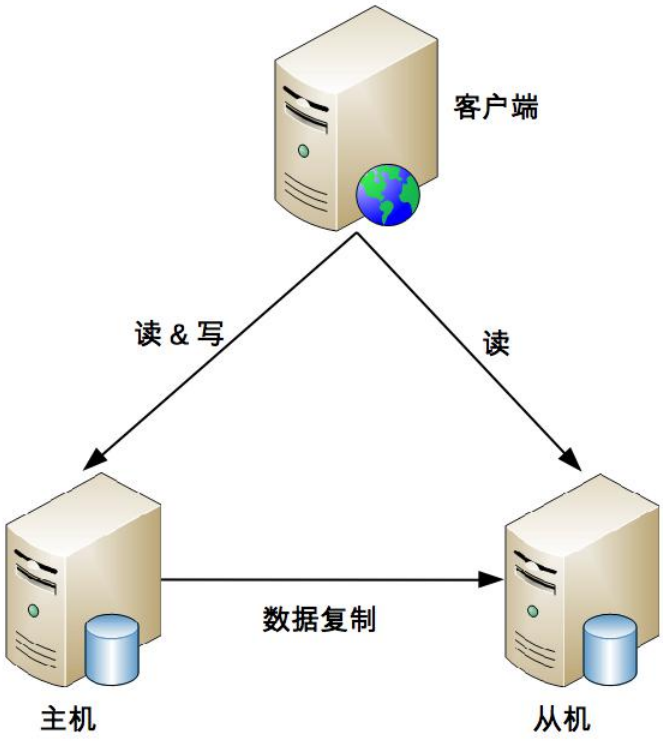
综合主备复制架构的优缺点，内部的后台管理系统使用主备复制架构的情况会比较多，例如学生管理系统、员工管理系统、假期管理系统等，因为这类系统的数据变更频率低，即使在某些场景下丢失数据，也可以通过人工的方式补全。

## 主从复制

主从复制和主备复制只有一字之差，“从”意思是“随从、仆从”，“备”的意思是备份。我们可以理解为仆从是要帮主人干活的，这里的干活就是承担“读”的操作。也就是说，主机负责读写操作，从机只负责读操作，不负责写操作。

1. 基本实现

下面是标准的主从复制架构：




与主备复制架构比较类似，主要的差别点在于从机正常情况下也是要提供读的操作。

2. 优缺点分析

主从复制与主备复制相比，优点有：

- 主从复制在主机故障时，读操作相关的业务可以继续运行。
- 主从复制架构的从机提供读操作，发挥了硬件的性能。

缺点有：

- 主从复制架构中，客户端需要感知主从关系，并将不同的操作发给不同的机器进行处理，复杂度比主备复制要高。
- 主从复制架构中，从机提供读业务，如果主从复制延迟比较大，业务会因为数据不一致出现问题。
- 故障时需要人工干预。

综合主从复制的优缺点，一般情况下，写少读多的业务使用主从复制的存储架构比较多。例如，论坛、BBS、新闻网站这类业务，此类业务的读操作数量是写操作数量的 10 倍甚至 100 倍以上。

## 双机切换

1. 设计关键

主备复制和主从复制方案存在两个共性的问题：

- 主机故障后，无法进行写操作。
- 如果主机无法恢复，需要人工指定新的主机角色。

双机切换就是为了解决这两个问题而产生的，包括主备切换和主从切换两种方案。简单来说，这两个方案就是在原有方案的基础上增加“切换”功能，即系统自动决定主机角色，并完成角色切换。由于主备切换和主从切换在切换的设计上没有差别，我接下来以主备切换为例，一起来看看双机切换架构是如何实现的。

要实现一个完善的切换方案，必须考虑这几个关键的设计点：

- 主备间状态判断

主要包括两方面：状态传递的渠道，以及状态检测的内容。

状态传递的渠道：是相互间互相连接，还是第三方仲裁？

状态检测的内容：例如机器是否掉电、进程是否存在、响应是否缓慢等。

- 切换决策

主要包括几方面：切换时机、切换策略、自动程度。

切换时机：什么情况下备机应该升级为主机？是机器掉电后备机才升级，还是主机上的进程不存在就升级，还是主机响应时间超过 2 秒就升级，还是 3 分钟内主机连续重启 3 次就升级等。

切换策略：原来的主机故障恢复后，要再次切换，确保原来的主机继续做主机，还是原来的主机故障恢复后自动成为新的备机？

自动程度：切换是完全自动的，还是半自动的？例如，系统判断当前需要切换，但需要人工做最终的确认操作（例如，单击一下“切换”按钮）。

- 数据冲突解决

当原有故障的主机恢复后，新旧主机之间可能存在数据冲突。例如，用户在旧主机上新增了一条 ID 为 100 的数据，这个数据还没有复制到旧的备机，此时发生了切换，旧的备机升级为新的主机，用户又在新的主机上新增了一条 ID 为 100 的数据，当旧的故障主机恢复后，这两条 ID 都为 100 的数据，应该怎么处理？

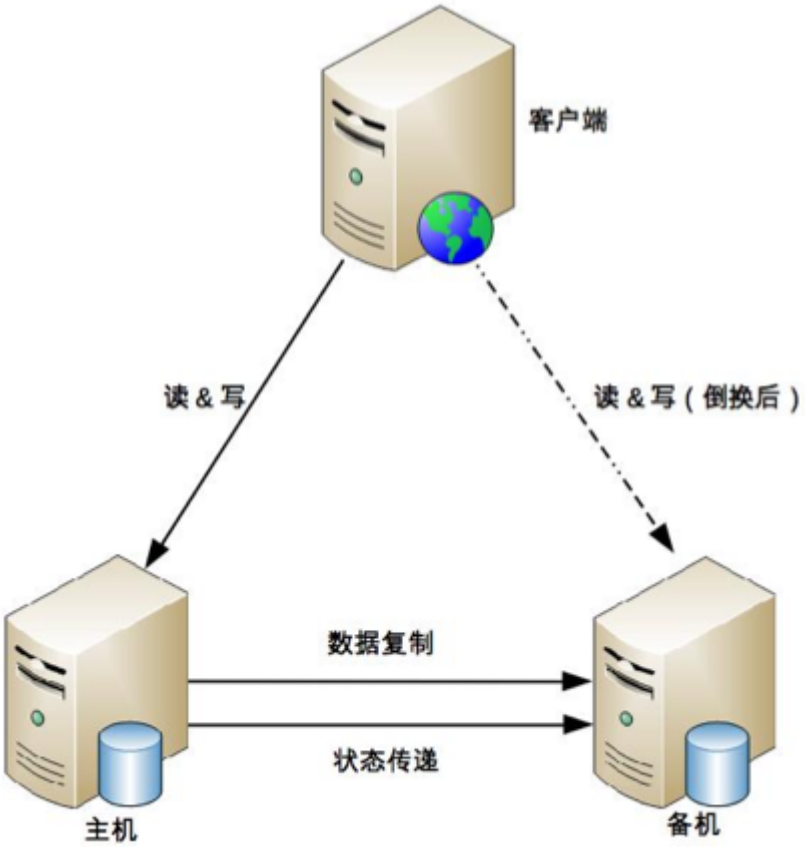
以上设计点并没有放之四海而皆准的答案，不同的业务要求不一样，所以切换方案比复制方案不只是多了一个切换功能那么简单，而是复杂度上升了一个量级。形象点来说，如果复制方案的代码是 1000 行，那么切换方案的代码可能就是 10000 行，多出来的那 9000 行就是用于实现上面我所讲的 3 个设计点的。

2. 常见架构

根据状态传递渠道的不同，常见的主备切换架构有三种形式：互连式、中介式和模拟式。

互连式

故名思议，互连式就是指主备机直接建立状态传递的渠道，架构图请注意与主备复制架构对比。




你可以看到，在主备复制的架构基础上，主机和备机多了一个“状态传递”的通道，这个通道就是用来传递状态信息的。这个通道的具体实现可以有很多方式：

- 可以是网络连接（例如，各开一个端口），也可以是非网络连接（用串口线连接）。
- 可以是主机发送状态给备机，也可以是备机到主机来获取状态信息。
- 可以和数据复制通道共用，也可以独立一条通道。
- 状态传递通道可以是一条，也可以是多条，还可以是不同类型的通道混合（例如，网络 + 串口）。

为了充分利用切换方案能够自动决定主机这个优势，客户端这里也会有一些相应的改变，常见的方式有：

- 为了切换后不影响客户端的访问，主机和备机之间共享一个对客户端来说唯一的地址。例如虚拟 IP，主机需要绑定这个虚拟的 IP。
- 客户端同时记录主备机的地址，哪个能访问就访问哪个；备机虽然能收到客户端的操作请求，但是会直接拒绝，拒绝的原因就是“备机不对外提供服务”。

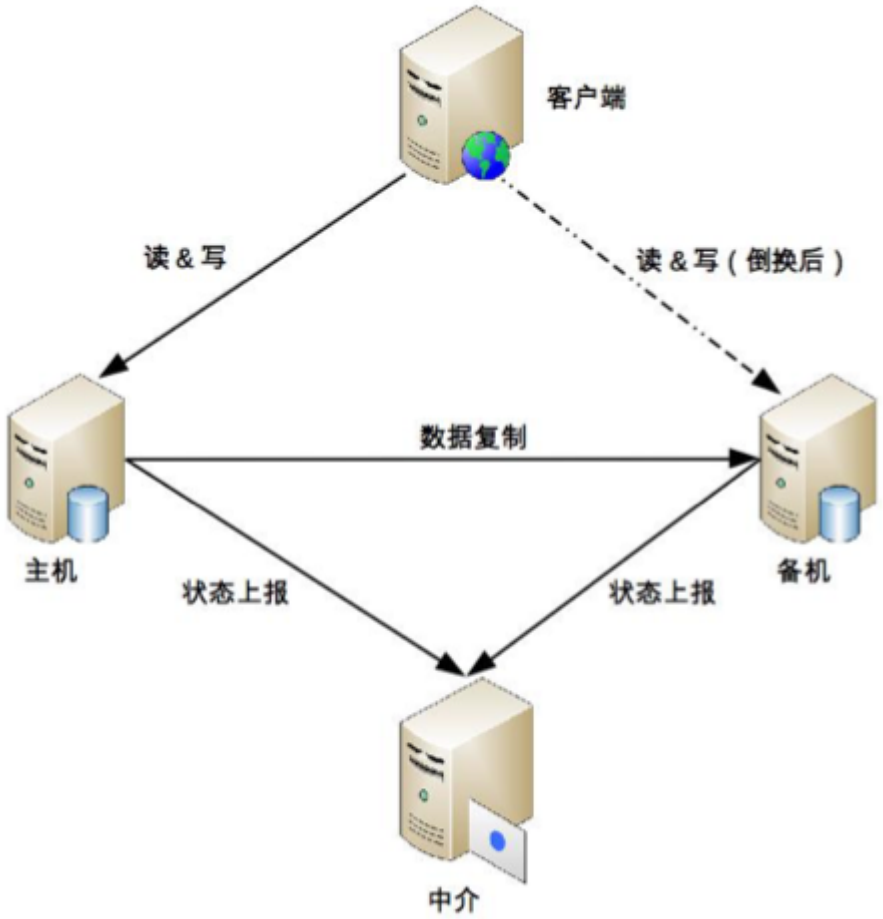
互连式主备切换主要的缺点在于：

- 如果状态传递的通道本身有故障（例如，网线被人不小心踢掉了），那么备机也会认为主机故障了从而将自己升级为主机，而此时主机并没有故障，最终就可能出现两个主机。

- 虽然可以通过增加多个通道来增强状态传递的可靠性，但这样做只是降低了通道故障概率而已，不能从根本上解决这个缺点，而且通道越多，后续的状态决策会更加复杂，因为对备机来说，可能从不同的通道收到了不同甚至矛盾的状态信息。

### 中介式

中介式指的是在主备两者之外引入第三方中介，主备机之间不直接连接，而都去连接中介，并且通过中介来传递状态信息，其架构图如下：




对比一下互连式切换架构，我们可以看到，主机和备机不再通过互联通道传递状态信息，而是都将状态上报给中介这一角色。单纯从架构上看，中介式似乎比互连式更加复杂了，首先要引入中介，然后要各自上报状态。然而事实上，中介式架构在状态传递和决策上却更加简单了，这是为何呢？

连接管理更简单：主备机无须再建立和管理多种类型的状态传递连接通道，只要连接到中介即可，实际上是降低了主备机的连接管理复杂度。

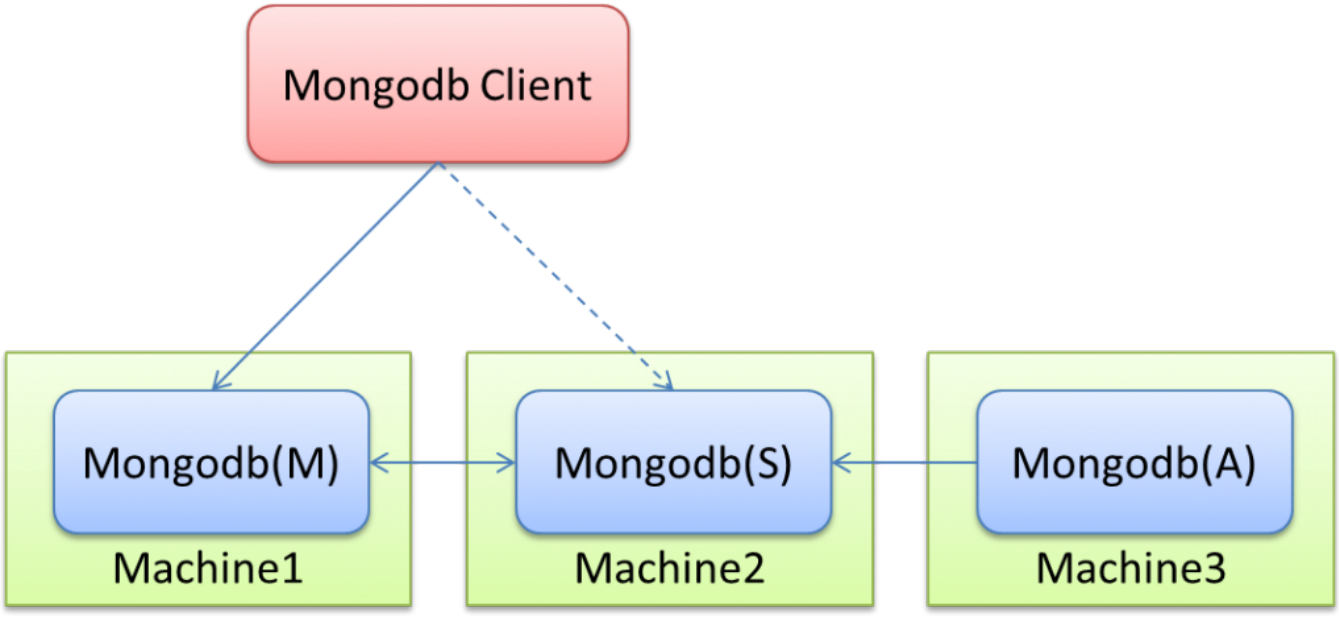
例如，互连式要求主机开一个监听端口，备机来获取状态信息；或者要求备机开一个监听端口，主机推送状态信息到备机；如果还采用了串口连接，则需要增加串口连接管理和数据读取。采用中介式后，主备机都只需要把状态信息发送给中介，或者从中介获取对方的状态信息。无论是发送还是获取，主备机都是作为中介的客户端去操作，复杂度会降低。

状态决策更简单：主备机的状态决策简单了，无须考虑多种类型的连接通道获取的状态信息如何决策的问题，只需要按照下面简单的算法即可完成状态决策。

- 无论是主机还是备机，初始状态都是备机，并且只要与中介断开连接，就将自己降级为备机，因此可能出现双备机的情况。
- 主机与中介断连后，中介能够立刻告知备机，备机将自己升级为主机。
- 如果是网络中断导致主机与中介断连，主机自己会降级为备机，网络恢复后，旧的主机以新的备机身份向中介上报自己的状态。
- 如果是掉电重启或者进程重启，旧的主机初始状态为备机，与中介恢复连接后，发现已经有主机了，保持自己备机状态不变。
- 主备机与中介连接都正常的情况下，按照实际的状态决定是否进行切换。例如，主机响应时间超过 3 秒就进行切换，主机降级为备机，备机升级为主机即可。

虽然中介式架构在状态传递和状态决策上更加简单，但并不意味着这种优点是没有代价的，其关键代价就在于如何实现中介本身的高可用。如果中介自己宕机了，整个系统就进入了双备的状态，写操作相关的业务就不可用了。这就陷入了一个递归的陷阱：为了实现高可用，我们引入中介，但中介本身又要求高可用，于是又要设计中介的高可用方案……如此递归下去就无穷无尽了。

MongoDB 的 Replica Set 采取的就是这种方式，其基本架构如下：



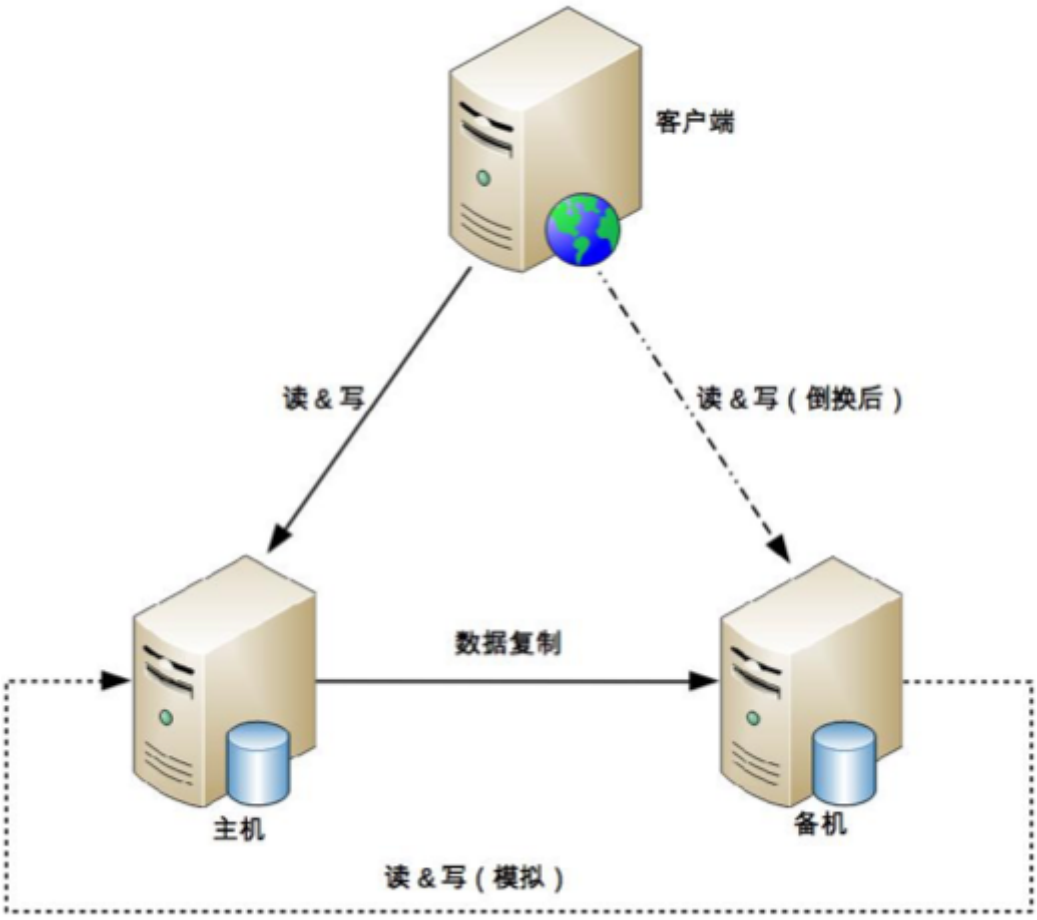

(http://img.my.csdn.net/uploads/201301/13/1358056331_2790.png)

MongoDB(M) 表示主节点，MongoDB(S) 表示备节点，MongoDB(A) 表示仲裁节点。主备节点存储数据，仲裁节点不存储数据。客户端同时连接主节点与备节点，不连接仲裁节点。

幸运的是，开源方案已经有比较成熟的中介式解决方案，例如 ZooKeeper 和 Keepalived。ZooKeeper 本身已经实现了高可用集群架构，因此已经帮我们解决了中介本身的可靠性问题，在工程实践中推荐基于 ZooKeeper 搭建中介式切换架构。

### 模拟式

模拟式指主备机之间并不传递任何状态数据，而是备机模拟成一个客户端，向主机发起模拟的读写操作，根据读写操作的响应情况来判断主机的状态。其基本架构如下：




对比一下互连式切换架构，我们可以看到，主备机之间只有数据复制通道，而没有状态传递通道，备机通过模拟的读写操作来探测主机的状态，然后根据读写操作的响应情况来进行状态决策。

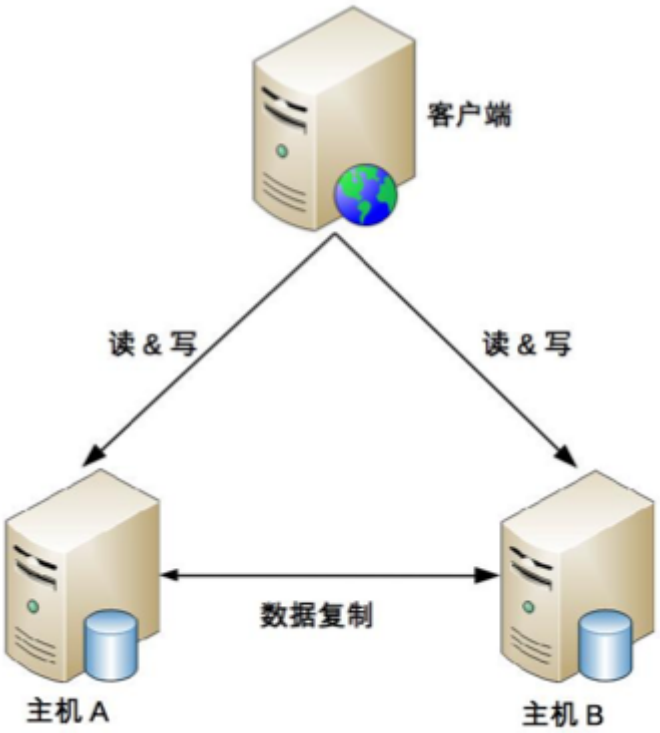
模拟式切换与互连式切换相比，优点是实现更加简单，因为省去了状态传递通道的建立和管理工作。

简单既是优点，同时也是缺点。因为模拟式读写操作获取的状态信息只有响应信息（例如，HTTP 404，超时、响应时间超过 3 秒等），没有互连式那样多样（除了响应信息，还可以包含

CPU 负载、I/O 负载、吞吐量、响应时间等），基于有限的状态来做状态决策，可能出现偏差。

## 主主复制

主主复制指的是两台机器都是主机，互相将数据复制给对方，客户端可以任意挑选其中一台机器进行读写操作，下面是基本架构图。




相比主备切换架构，主主复制架构具有如下特点：

- 两台都是主机，不存在切换的概念。
- 客户端无须区分不同角色的主机，随便将读写操作发送给哪台主机都可以。

从上面的描述来看，主主复制架构从总体上来看要简单很多，无须状态信息传递，也无须状态决策和状态切换。然而事实上主主复制架构也并不简单，而是有其独特的复杂性，具体表现在：如果采取主主复制架构，必须保证数据能够双向复制，而很多数据是不能双向复制的。例如：

- 用户注册后生成的用户 ID，如果按照数字增长，那就不能双向复制，否则就会出现 X 用户在主机 A 注册，分配的用户 ID 是 100，同时 Y 用户在主机 B 注册，分配的用户 ID 也是 100，这就出现了冲突。
- 库存不能双向复制。例如，一件商品库存 100 件，主机 A 上减了 1 件变成 99，主机 B 上减了 2 件变成 98，然后主机 A 将库存 99 复制到主机 B，主机 B 原有的库存 98 被覆盖，变成了 99，而实际上此时真正的库存是 97。类似的还有余额数据。

因此，主主复制架构对数据的设计有严格的要求，一般适合于那些临时性、可丢失、可覆盖的数据场景。例如，用户登录产生的 session 数据（可以重新登录生成）、用户行为的日志数据（可以丢失）、论坛的草稿数据（可以丢失）等。

## 小结

今天我为你讲了高可用存储架构中常见的双机架构，分析了每类架构的优缺点以及适应场景，希望对你有所帮助。

这就是今天的全部内容，留一道思考题给你吧，如果你来设计一个政府信息公开网站的信息存储系统，你会采取哪种架构？谈谈你的分析和理由。

欢迎你把答案写到留言区，和我一起讨论。相信经过深度思考的回答，也会让你对知识的理解更加深刻。（编辑乱入：精彩的留言有机会获得丰厚福利哦！）




精选留言

空档滑行 10

政府信息网站使用主备或者主从架构就可以了。信息都是人工录入，可以补录。数据本来对实时性要求不高，所以出了故障人工修复也来得及。所以主备就够了，如果为了照顾形象可以用主从，保证主机故障后仍然可以查，不能新发

2018-06-23

作者回复

分析正确

2018-06-23

今夕是何年 3

政府信息网站使用主从就行了，因为读的请求多，写的请求少。
网站挂掉影响也不大，所以可以不用主从切换。

2018-06-23

南友力max先森 1

单机就可以了，搞那么复杂

2018-07-16

作者回复

单机可靠性只有2个9

2018-07-16

忠厚 1

数据持久化信息我可能会选择主备模式，备机主做数据备份不提供读写操作。

添加一个redis缓存全量信息数据，做一个哨兵模式，实现故障切换，提高网站的可用性

应用上再使用个Ehcache堆外缓存，主要把热点信息放到应用里提升性能.

这样做相对主从模式，读并发压力过大时，扩容更容易

2018-06-27

作者回复

缓存设计得比较复杂了，我认为ehcache没有必要

2018-06-28



gen_jin 1

我认为对政府信息系统：
1. 由于数据写少读多（1：10000）：采用主从复制（利用从机读）而不是主备。
2.由于面对公众性，最好24小时无间断工作，出现故障最好采用自动双机切换；而考虑将来扩容，开始是一主一从 后面是一主多从，对一主多从 实现简单看最好中介式（使用zk或LVS+Keepalived的架构 实现一主多从）。
李老师，以上仅是我的片面之词，欢迎多多指教！

2018-06-23

作者回复

用了主从复制即可，没必要切换，因为写很少

2018-06-25



A:春哥大魔王 0

华哥你好，您作为资深技术专家方便总结下成为技术专家，架构师等高级技术职位所需要进阶的技术能力和软技能吗。非常期待

2018-08-06

作者回复

整理了架构师技能图谱，等编辑发布😀

2018-08-08

孙振超 0

双机切换架构里面的中介模式是由db连接到中介，而后中介告诉它应该是主还是备，这种模式下要求db能够根据中介的返回结果实时的修改自己的模式，同时当客户端请求类型和当前模式不匹配时返回调用失败。对于mongo db原生支持还是可以，如果是原生不支持的db，是不是改为客户端直接链接中介，根据请求类型获取对应的db ip可用性更好些，如同zookeeper？又或者mongo db采用客户端直接链接中介是否也可以？因为中介模式本身对中介的高可用要求也比较高。

2018-08-06

作者回复

客户端直连中介，需要中介理解存储系统的协议，这个做不到通用，MySQL Router可以实现你说的功能，但只适应MySQL，如果你基于zk做一套中介，可以支撑MySQL, mongodb等

2018-08-08

fiseasky 0

如何做主备复制，比如redis, mongdb等？只是简单的配置还是需要单独写代码来实现呢？

2018-07-30

作者回复

基本都是配置就可以用了

2018-07-30

listen to you 0

请问主从数据一致性如何保证？如何补救？

2018-07-27

作者回复

没法保证实时一致性，最终一致性依赖存储系统的同步就可以了

2018-07-30

Geek_8242cb 0

老师好！文件服务器同城双中心高可用有什么好的开源实现方案吗？nfs据说有安全问题，还有别的好方案吗？

2018-07-18

作者回复

我们用过ocfs，但现在已经切换阿里云了，开源方案试试ceph或者Moose File System。

另外，咨询了我们的运维大神，内网NFS不用担心安全问题

2018-07-19

小飞哥 超級會員 0

我想说主主复制和摸拟式的也会有高可用的问题啊吧！多个备机去主机模拟写若不通则谁来当主机的问题

2018-07-17

作者回复

主主只有两台机器，多台备机那就是集群了

2018-07-18

joyafa 0

政府网站，访问量也不会太大，基本上就一个留言窗口需要写数据，其他地方都是读操作，写很少，一主一备就好了。

2018-07-11

作者回复

主从更好一些

2018-07-14



100kg 0

随着用户量的增大，肯定要上多主多备的，这样的情况对于“库存”字段该怎么处理呢?

2018-07-10

作者回复

考虑后面介绍的数据分散集群，集群中每个机器存储一部分库存数据

2018-07-11

枫晴 andy 0

政府公布信息应该是读多写少的场景，使用主从架构做读写分离保证高可用。
另外，有个问题，主主架构到底能用在哪些场景不是特别清楚，感觉这种必须是两边都能读写的数据必须分片？否则两边都写会冲突。

2018-07-08

作者回复

主主应用比较少，一般用在管理或者维护系统，数据冲突影响不大，或者设计比较复杂的数据生成方案，例如A机房只会生成奇数数据，B机房只会生成偶数数据

2018-07-09



100kg 0

那老师，如果采用主主的方式做mysql集群，对于库存字段该怎么处理呢？ 乐观锁管用吗?

2018-07-07

作者回复

库存别用主主架构

2018-07-09

王虹凯 0

关于主主互相复制相关的有没有进一步讲解。

2018-07-06

作者回复

没太多可讲的呢，主要是设计数据防冲突策略和冲突解决方案，例如A机房生成奇数数据，B机房生成偶数数据

2018-07-09

来 0

楼主你好，我看你文章中也提到了主从复制延迟和复制中断，我想问下当你们系统出现复制延迟后你们采用的处理方案是什么，还有就是复制中断（可能是网络闪断，也可能是从机宕掉，从机再次启动）后，当主从再次连接后，主机是如何准确的把未同步的数据同步到这台从机上的呢，因为我一个主机可能有多个从机，只是其中一个从机有问题，麻烦楼主帮忙解答下，谢谢

2018-07-05

作者回复

1. 复制延迟一般只能等

2. 复制中断后的恢复，需要有机制判断复制进度和位置，参考mysql的binlog复制

2018-07-05



jacy 0

政务网的站读写量差距巨大，对读稳定有要求，出但出问题也不至于要求马上恢复，可以选择主从，然后人工恢复即可。

2018-07-04

作者回复

赞同

2018-07-04

蛤蟆不好找 0

对比主备主从，优势是主从更适合，继续政府的网站的信息，有的时候可能会因为某个政策大量的涌入读取，这个时候主从的优势就很明显的展露了，因为有从机，所以可以分流部分的QPS，降低主机的压力

2018-07-02

作者回复

用缓存更好😄

2018-07-02

来 0

采用主从即可，政府信息公开网站读写压力都比较小，大部分是读，即使写库宕掉也没关系，写数据丢失了直接补录进去就行。主宕掉直接人工切换主备就行，没必要太复杂

2018-07-01

作者回复

赞同👍

2018-07-02