

# 分布式关系数据库 OceanBase 的高可靠性

文 || 蚂蚁金融服务集团高级研究员 阳振坤



阳振坤，在北京大学先后获得数学学士、硕士及计算机博士学位后留校。1999 年成为北京大学首批“长江学者奖励计划”特聘教授。先后获得北京市科学技术进步奖一等奖、国家科学技术进步奖一等奖（排名第四）、第六届中国青年科技奖等。目前是阿里巴巴 / 蚂蚁金服的 OceanBase 关系数据库项目负责人。

**数**据库是现代金融系统最关键的基础设施之一，迄今为止，主要商业数据库本质上是单机系统，其容量、性能和可靠性均依赖于单个（少量）高性能服务器与单个高可靠存储的组合，成本高昂且扩展困难。阿里巴巴集团和蚂蚁金融服务集团自研的 OceanBase 分布式关系数据库采用 PC 服务器机群，不仅易于扩展，还通过多库多活和分布式选举等技术，以廉价 PC 服务器机群获得了极高的可靠性，并在阿里巴巴集团和蚂蚁金融服务集团得到了广泛应用，经受了多次“双十一”的考验。

## 一、背景

网上购物的迅猛发展和移动支付的飞速普及不断创造着生产系统数据库事务处理性能的新记录，例如在 2015 年 11 月 11 日（天猫“双十一”）凌晨的 00:05:01 和 00:09:02，中国与世界各地消费者一起在支付宝分别创下了订单创建 14 万笔 / 秒和订单支付 8.59 万笔 / 秒的记录。传统商业数据库的扩展能力面临着严峻的挑战。

OceanBase 是一个分布式的关系数据库系统，具备传统关系数据库的功能，包括标准 SQL 语言支持、事务 (ACID)、二级索引、范围查询、表的连接等。与传统关系数据库相比，OceanBase 有如下的不同。

首先，OceanBase 以 PC 服务器机群代替了传统关系数据库的单个（少量）高性能服务器。PC 服务器的突出优点是性价比非常高、采购周期短、机群易于扩容。但与高性能服务器相比，PC 服务器的单机处理能力和可靠性都相对较低。

其次，OceanBase 以 PC 服务器自带硬盘（机械盘或者固态硬盘）代替了传统关系数据库的高可靠存储。PC 服务器自带硬盘的突出优点是性价比高、采购周期短。但与高可靠存储相比，PC 服务器单机的自带硬盘容量小、可靠性低，一旦服务器或者硬盘故障，则其上的数据会暂时不可用甚至彻底丢失。

相比于传统关系数据库，OceanBase 的扩展能力和性价比优势十分突出。与此同时，PC 服务器及其硬盘的高性价比与相对低的可靠性则是一对孪生兄弟：PC 服务器及其硬盘大约只有 99.9% 左右的可靠性，远远低于核心业务的 99.999% 的可靠性需求。为了服务商业和金融等业务，OceanBase 通过多库多活、自动容灾、版本灰度升级和远程灾备等技术在不可靠的 PC 服务器及其硬盘的基础上实现了非常高的可靠性。

## 二、多库多活

传统关系数据库通过主备镜像进一步提升数据库的可靠性，但是，除非使用高可靠的、主库备库共享的存储，否则无法保证主库在故障发生的时候能够把最新提交的事务全部同步到备库。OceanBase 采用了多库多活的方式（如图 1 所示），多个库（通常

$\geq 3$ ) 中, 一个是主库, 其余是备库, 主库执行事务并实时同步给备库: 仅当事务同步到包括主库在内的多数库(例如 3 个库中至少 2 个, 5 个库中至少 3 个等)并持久化后, 事务才成功。

任一时刻, 如果少数库故障, 例如 3 个库中的 1 个, 假如是 1 个备库故障, 则业务不会受到任何影响; 假如是主库故障, 由于已经成功的每一笔事务到达了至少一个备库, 在两个备库选举出新的主库并进行必要的相互同步后, 即可继续提供服务, 不会出现任何数据丢失。

单台 PC 服务器的可靠性较低, 但两台 PC 服务器同时出故障的概率极低, 因此 3 机群部署时, 除非两个机群的两台服务器同时故障, 否则不会有任何数据损失、业务也基本不受影响。

### 三、自动容灾

OceanBase 的主库是由几个库(例如 3 个库或者 5 个库等)根据 Paxos 协议和 DBA 指定的规则(例如距离应用服务器最近的库优先)选举出来的, 只有获得超过半数的选票才能成为主库。主库当选后只有有限的任期(例如几十秒)并通知所有备库其主库角色和任期。当前主库通常在任期结束前发起连任选举以求连任, 一般情况下主库的连任选举都是成功通过并通知所有备库其新的任期。如果连任选举失败, 则主库在自身任期结束时放弃主库身份, 新的主库随后将被选举出来。当前主库也可以根据需要发起改选以便使得另外一个库成为主库(例如当前主库所在机群即将进行版本升级或者当前主库所在服务器即将下线等)。

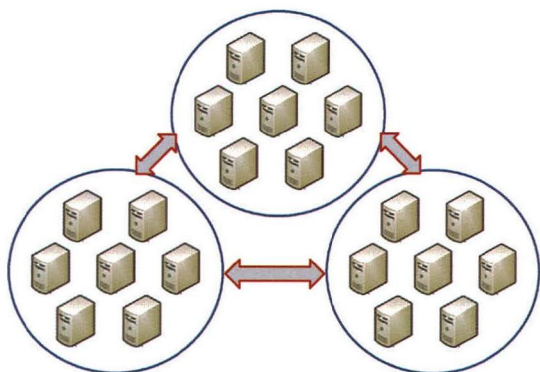


图1 OceanBase 的三机群部署(三库)

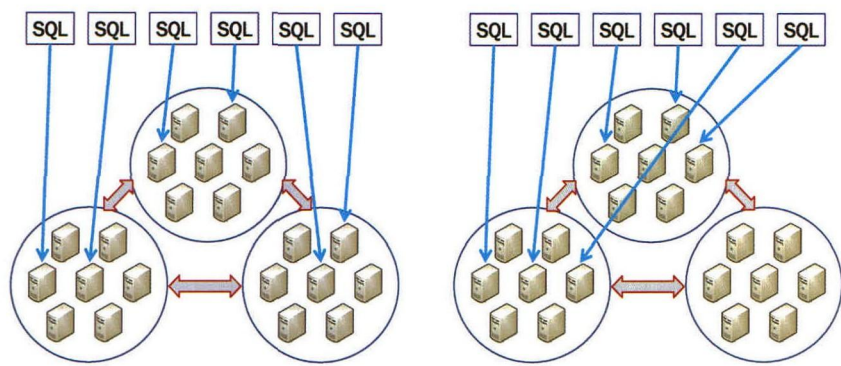
备库持有当前主库的地址和任期等信息, 如果备库持有的主库任期过期, 则备库认为当前主库不存在并自主地发起主库选举。如果主库崩溃或者主库网络中断, 则每个备库持有的主库任期都会过期, 这些备库最终都会进行选举并产生新的主库。如果只有少数备库的网络异常导致无法接收到主库的任期广播, 这些备库认为当前主库不存在, 它们会自主发起选举, 但它们无法成为多数, 不会影响当前主库。

如果由于某些极端原因导致多数库故障(例如电源或者网络中断等), 则系统因为无法自动选举出主库而报警。只要有一个库还存在, 则系统依然可以提供服务。如果存在的是主库, 则没有任何数据损失, 如果存在的是备库, 此时可能有数据损失, 由于备库落后于主库的时间低于几秒, 因此恢复点目标 RPO (Recovery Point Objective) 是几秒。

少数备库故障不会丢失数据, 也不影响业务。如果主库突然故障, 备库必须等待主库的任期过期后才能选举出新的主库并提供服务, 这个时间最长几十秒(略长于主库的任期), 即恢复时间目标 RTO (Recovery Time Objective) 最长几十秒(RPO 则为 0)。这段时间内, 部分数据服务不可用。假如机群有 20 台服务器, 那么单个服务器故障导致大约 5% 的数据服务受到影响。针对这种情况, 一些辅助措施可以采用, 例如硬盘是最容易损坏的部件, 对硬盘做 RAID 5 可以消除单块硬盘故障对业务的影响。再如定期对服务器(包括硬盘等)进行健康检查, 对“亚健康”或者检查异常的服务器上的主库进行主动的主备切换以便下线进行彻底检查, 而主动的主备库切换不会影响业务, 更不会丢失数据。以上措施使得服务器的突然故障十分罕见, 减少了主库的突然崩溃对业务的影响, 即使这个影响相对有限(几十秒内部分数据不可用且没有任何数据丢失或损坏)。

### 四、版本灰度升级

软件版本升级常常伴随着或多或少的风险, 数据库的版本升级(包括打补丁)更是如此, 包括数据兼容性与正确性问题、性能异常、功能变化等等。这些问题可能导致数据异常或者服务停顿, 对业务产生



各种程度甚至非常严重的影响。银行、企业和政府都不乏因为数据库版本升级（打补丁）导致的各种故障。为了避免版本升级导致的各种异常，OceanBase 利用自身的多库多活的特性，实现了自身软件版本的灰度升级（如图 2 所示）。

通常情况下，OceanBase 三机群部署（三库）下，每个机群内都有部分主库和部分备库，每个机群都有来自业务的访问流量。

OceanBase 的灰度升级每次升级一个机群：（1）切换待升级机群的主库到其余两个机群（切走写流量）；（2）切走该机群的读流量；（3）对该机群进行升级（打补丁）；（4）监控观察升级后的版本兼容性（例如日志同步及回放等）和机群健康状况，例如 CPU 负载、硬盘 IO、网络带宽占用等状况，正常则转到下一步；（5）进行内置的数据对比，完全正确则转到下一步；（6）逐步引入部分读流量，例如白名单、1% 流量等，如果正常，逐渐增加流量并继续上述步骤 4 和 5 的监控观察和数据对比；（7）引入部分写流量，例如白名单、1% 流量等，如果正常，逐渐增加流量并继续上述步骤 4 和 5 的监控观察和数据对比；（8）继续观察以确认新版本符合预期。

在以上的升级步骤中，任何一步出现异常都可即刻回退读写流量（如果有的话）至旧版本，确保了即使

图2 OceanBase 的灰度升级



出现异常，新版本对业务的影响是可控的。同样的方法可以升级另外的机群。


## 五、远程灾备

水灾、火灾、地震或其他灾害等可能导致一个地区的数据中心无法使用，因此 OceanBase 还提供了远程灾备。数据库的快照定期（例如每周一次）复制到远端，事务日志则准实时地复制到远端，因此远端的数据至多落后当前数据几十秒（除非通信链路中断），即能够保证 RPO 在几十秒之内。

恢复时间目标 RTO 则取决于远端灾备是热备还是冷备，如果是热备（成本较高），则 RTO 不超过几分钟，如果是冷备（成本较低），则 RTO 可能达若干小时。

## 六、小结

OceanBase 采用 PC 服务器机群代替可靠性很高同时也非常昂贵的高性能服务器和高可靠存储，获得了很高的性价比。为了克服 PC 服务器及其硬盘相对较低的可靠性对业务的影响，OceanBase 通过多库多活的方式保证了数据库数据不会因为少量服务器或硬盘故障而不可访问或丢失；通过 Paxos 协议自动选举主库保证了服务的连续性；通过灰度升级避免了版本升级带来的灾难和数据错误；通过远程灾备避免了水灾火灾地震等灾害导致的数据中心不可用而引起的服务中断。

OceanBase 已经在阿里巴巴集团和蚂蚁金融服务集团内广泛使用，连续五年参加了天猫“双十一”大促，并在 2015 年“双十一”支撑了支付宝的全部国内和国际交易，创造了 14 万笔/秒的交易订单创建的记录。

### 参考资料：

- [1] Daniel Peng and Frank Dabek, Large-scale Incremental Processing Using Distributed Transactions and Notifications, Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation, USENIX (2010)
- [2] Irving L. Traiger, James N. Gray, Cesare A. Galtieri, Bruce G. Lindsay, Transactions and Consistency in Distributed Database Systems, ACM Transactions on Database Systems (TODS), Volume 7 Issue 3, Sept. 1982, Pages 323-342
- [3] James C. Corbett, Jeffrey Dean, etc., Spanner: Google's Globally-Distributed Database, In Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation (Berkeley, CA, USA, 2012), OSDI'12, pp. 251-264