CAICT 中国信通院　　Bai 百度

金融级数据库容灾技术报告

（2021 年）

中国信息通信研究院云计算与大数据研究所
北京百度网讯科技有限公司

2021 年 7 月

版权声明

前　言

数据库作为信息系统的核心，承担着数据存储和业务逻辑计算等工作，各金融机构信息化部门均应用大量容灾备份手段，用于保障数据库的业务连续性。近年来，随着数据库应用场景日益丰富和数据库的云化趋势显著，金融机构数据库的容灾备份手段和模式也不断迭代更新。

本报告聚焦金融领域的数据库在灾备方面的技术内容。介绍了容灾与备份的定义、分类，分析了金融机构灾备现状、需求与灾备市场情况，梳理了主流数据库容灾备份技术架构、实现方式与部署方案，阐述了节点、网络、数据中心等故障的应对方案，展望了数据库灾备技术发展方向。由于时间仓促，水平所限，错误和不足之处在所难免，欢迎各位读者批评指正，意见建议请发送至liusiyuan@caict.ac.cn。

目 录

图 目 录

表 目 录

一、容灾备份概述

容灾备份简称“灾备”，是指利用科学的技术手段和方法，提前建立系统化的数据应急方式，以应对灾难的发生。**容灾与备份是两个独立的概念，容灾是为了在遭遇灾害时保证信息系统能正常运行，帮助企业实现业务连续性的目标，备份是为了应对灾难来临时造成的数据丢失问题。**

(一) 容灾备份基础介绍

1.灾备由来及定义

灾备行业起源于20世纪70年代的美国费城。1979年，SunGard公司在费城建立了全世界第一个灾备中心，当时人们关注的重点是企业的数据备份和系统备份。后来，IT备份发展到了灾难恢复规划（DRP），在IT备份中加入了灾难恢复预案、资源需求、灾备中心管理，形成了对生产运行中心的保障概念。再后来，人们把灾难恢复从IT角度逐渐转向了业务的角度，用业务来衡量灾备目标，即哪些业务最重要，哪些业务可容忍的恢复时间最短。随着企业规模扩展及信息系统的应用范围日益扩大，信息系统在企业运营过程中的角色愈发重要，为防范因为各种因素企业数据遭到毁坏，如地震、火灾、恐怖袭击等，异地灾备建设的需求应运而生。

2.灾备发生方式

根据已经发生的灾备事件总结分析，通常情况下灾难发生的原因有以下几种方式:

表 1 信息灾备发生方式

灾难原因	具体方式	典型事件
IT系统问题	服务器、应用程序、数据库等故障	2017 年 1 月 Gitlab 遭受 DDoS攻击，导致数据库写入锁定，网站出现不稳定和宕机。
信息安全管理问题	升级失败、权限管理混乱	2018年4月，英国TSB银行系统升级严重混乱。
灾害类事件	地震、洪灾、火灾	2014年4月，三星位于首尔郊区的果川的数据中心发生火灾，导致三星服务宕机。

来源：公开资料整理

根据研究表明[1]，在以上各种导致系统故障的原因中，其占比分别为：约40%的系统故障由操作人员失误引起，约40%的系统故障由应用软件的问题所引起，约20%的系统故障由设备的物理原因引起，如硬件失效、掉电等。综上，系统故障导致业务下线的主要原因在于人为操作失误或应用软件问题，而自然灾难引发的业务下线概率较小。

3.容灾定义及分类

容灾，即灾难发生时，在保证生产系统数据尽量少丢失的情况下，保持生产系统业务的不间断运行。容灾技术是信息系统的高可用性技术的一个组成部分。**容灾方式根据容灾距离和保护等级存在两种分类方式。**

（1）按容灾距离划分

按照容灾距离分类，容灾可分为本地容灾和异地容灾。

本地容灾一般指主机集群，当某台主机发生故障，其它主机可以

[1]《2021 中国灾备行业白皮书》，英方研究院。

代替该主机，继续正常对外提供服务。通常可以通过共享存储或双机双柜的方式实现本地容灾，其中多以共享存储为主。共享存储由活动主节点、不活动备节点和共享存储三部分组成。其中两台计算资源节点提供主备角色服务，通过SAN网络附加型存储作为数据存储的介质。主备节点共享一份存储，一旦主节点宕机，备节点可基于共享存储实现业务的接管。共享存储的短板在于远距离高可用接管成本较高，存在较大存储故障风险，且只支持一对一架构。

异地容灾是指在与生产机房一定距离的异地建立与生产机房类似的备份中心，并采用特定的技术将生产中心的数据传输到备份中心。传统异地容灾通过磁盘或磁带备份手段，对本地关键数据进行备份，然后运输至生产中心之外的地方进行保存，灾难发生后，可通过磁盘或磁带实现系统和数据的恢复。这种手段成本低、易操作，但是当存储数据大规模增加时，存储介质管理将成为难以解决的问题。现多采用网络传输的方式进行异地容灾。

（2）按保护级别划分

按照保护级别，容灾系统可分为数据级容灾、应用级容灾和业务级容灾。

数据级容灾是最基础的手段，指通过建立异地容灾中心，进行数据的远程备份，发生灾难时应用会中断。这种级别的容灾方案实施相对简单、资源投入和后期运维成本低，但系统恢复速度较慢，业务恢复难度高。

应用级容灾主要针对关键应用进行的容灾方案，应用级容灾是建

立在数据级容灾基础上，对应用系统进行实时复制，即在备份站点构建一套相同的应用系统，通过同步或异步复制技术，保障关键应用在允许的时间范围内恢复运行。应用级容灾实施难度高、资源投入和后期运维成本较高，需要更多的软件实现，但是系统恢复速度较快，业务恢复难度较低。

业务级容灾是最高级别的容灾手段，它包括除了保障IT系统业务连续性外也提供非IT系统保障，业务级容灾是在数据级容灾和应用级容灾基础之上，还需要考虑IT系统之外的业务因素。如发生重大灾难时，用户的办公场所可能会被损坏，除了恢复原来的数据外，还需要工作人员在备份的工作场所能够正常地开展业务。

4.备份定义及分类

备份是指数据或系统的备份，它是容灾的基础，是指为防止系统出现操作失误或故障导致的数据丢失，而将全部或部分数据集合从应用主机的硬盘或阵列复制到其它存储介质的过程，数据库的备份与恢复通常基于数据库日志文件进行操作。**备份方式根据备份数据量、备份频率和备份对象等多种分类方式。**

按照备份数据量，备份可分为全量备份、增量备份和差异备份。

全量备份是指用存储介质对整个数据及系统进行完全备份。这种备份方式易理解、易恢复；短板是在备份数据中有大量的重复数据，由于需要备份的数据量相当大，因此备份时间较长。

增量备份是指备份自上一次备份（包含全量备份、差异备份、增量备份）之后有变化的数据。增量备份过程中，只备份有标记的选中

的文件和文件夹。这种备份的优点是重复数据少，即节省存储空间又缩短了备份时间。

差异备份是指备份上一次全量备份之后有变化的数据，差异备份后不标记为已备份文件，进行恢复时，只需对第一次全量备份和最后一次差异备份进行恢复。增量备份与差异备份的差异是，增量备份判断数据更新标准是依据上一次备份检查点，而差异备份一定是依据全量备份检查点。如没有全量备份，就没有差异备份。差异备份的主要目的是限制完全恢复时使用的介质数量。

表 2 备份类型及差异

备份类型	原理	优点	缺点
全量备份	对备份集合所有数据进行备份	完全恢复系统需要的时间最短	费时，如果文件不频繁更改，备份内容几乎完全相同
增量备份	对上次备份后改变的数据进行备份	存储的数据最少，备份速度最快	完全恢复系统需要时间比全量或差异备份长
差异备份	对自上次全量备份后改变的数据进行备份	恢复时仅需要最新全量备份和相应的差异备份，速度比全量备份快	完全恢复系统需要时间比全量备份长，如果大量数据发生变化，备份所需时间长于增量备份

来源：2021中国灾备行业白皮书，英方研究院

按照备份频率，备份可以分为定时备份和实时备份。定时备份是指有时间间隔的数据备份方式，比如一小时一次，一天一次，定时备份无法保证数据零丢失。**实时备份**是指无时间间隔的数据备份方式，通过实时数据复制，保证主备两端的数据读写一致，确保数据零丢失。

按照备份对象，备份可以分为字节级备份、块级备份和文件级备份。字节级备份是指以字节级变量为基本单位，通过捕获生产系统数据的变化，并将变化数据实时传输到备端。**块级备份**是指以磁盘块为

基本单位，将数据从源端复制到备端，即每次备份数据以一个扇区或多个连续扇区为单位来进行备份。**文件级备份**是指以文件为单位，将数据以文件的形式读出，通过文件系统接口调用备份到另一个介质上。此外，根据数据备份时服务器是否停机可分为冷备和热备，按照数据存储介质之间的距离又分为本地备份和异地备份。

5.容灾与备份的区别

备份是容灾的基石，其目的是为了系统数据崩溃时能够恢复数据。容灾不能替换备份，容灾系统会完整地将生产系统的任何变化复制到容灾端，比如误将计费系统内的用户信息表删除，容灾端的用户信息表也会被完整删除。如果是同步容灾，容灾端的相关数据同时被删除了；如果是异步容灾，容灾端的相关数据在数据异步复制的间隔内会被删除。这时需要从备份系统中取出最新备份，从而恢复被错误删除的信息。因此，容灾系统的建设不能替代备份系统的建设。

6.灾难恢复衡量指标

评估一个灾备系统可靠性的两个重要指标为恢复时间目标（Recovery Time Objective，以下简称“RTO”）与恢复点目标（Recovery Point Objective，以下简称“RPO”）。

RTO是恢复时间目标，指灾难发生后，从系统宕机导致业务停顿开始，到系统恢复至可以支持业务部门运作，业务恢复运营之时，此两点之间的时间，RTO可简单地描述为企业能容忍的恢复时间。根据金融行业标准规定[2]，金融领域分布式事务数据库灾难恢复能力应至

[2] JR/T 0205-2020《分布式数据库技术金融应用规范 灾难恢复要求》，中国人民银行。

少达到该标准中规定的4级以上，其中4级RTO≤30分钟、5级≤15分钟，6级≤1分钟。

RPO是恢复点目标，指灾难发生后，业务系统所能容忍的最大数据丢失量，它是衡量企业在发生灾难后会丢失多少生产数据的指标。根据上述金融行业标准，应用于金融领域的分布式事务数据库RP0要求均为0。

RTO与RPO的确定必须在进行风险分析和业务影响分析后根据不同的业务需求确定，不同企业统一业务，需求也会有所不同。最理想的情况是两个指标都为零。

(二)金融机构灾备现状与要求

1. 金融机构灾备现状与需求

金融机构对数据零丢失和业务连续性要求在各行各业要求最高。以银行为例，银行信息系统架构最为严格，要求采用两地三中心或主备等多种模式构建灾备系统，通过裸光纤或密集型光波复用（DWDN）技术实现数据中心与各个营业网点的数据同步。以证券公司为例，各大券商的灾备中心的架构通常采用虚拟化技术，实现生产端物理机与虚拟机并存，灾备端以虚拟机为主的配置[3]。

综合各类金融机构，主要需求包括但不限于：1）海量数据备份、实时复制；2）数据库数据跨平台迁移和读写分离；3）主备业务系统应用高可用；4）提升灾备中心智能运维水平；5）大规模灾备系统可用性验证的自动化能力等。

[3]《2021 中国灾备行业白皮书》，英方研究院。

2.金融机构容灾要求

金融机构分类较多，不同行业发布的法律法规各不相同，以要求最为严格的银行和证券行业为代表，以下列举相关重要的现行标准规范:

《证券期货业数据分类分级指引》自2018年9月27日同日公布实施，数据分类是按照GB/T 10113-2003中的线分类法和GB/T 22240-2008中的定级方法为基础进行分类的。目的是在数据分类的基础上，对已分类数据按照数据泄露或损坏造成的影响进行分级，形成统一的分类分级方法。同时，在数据用语的使用过程中，也强调予以统一。

除上述监管要求外，证券行业信息安全监管还有包括但不限于:

- 《证券期货业信息安全保障管理办法》;
- 《证券期货业信息安全事件报告与调查处理方法》;
- 《证券期货经营机构信息系统备份能力标准》等。

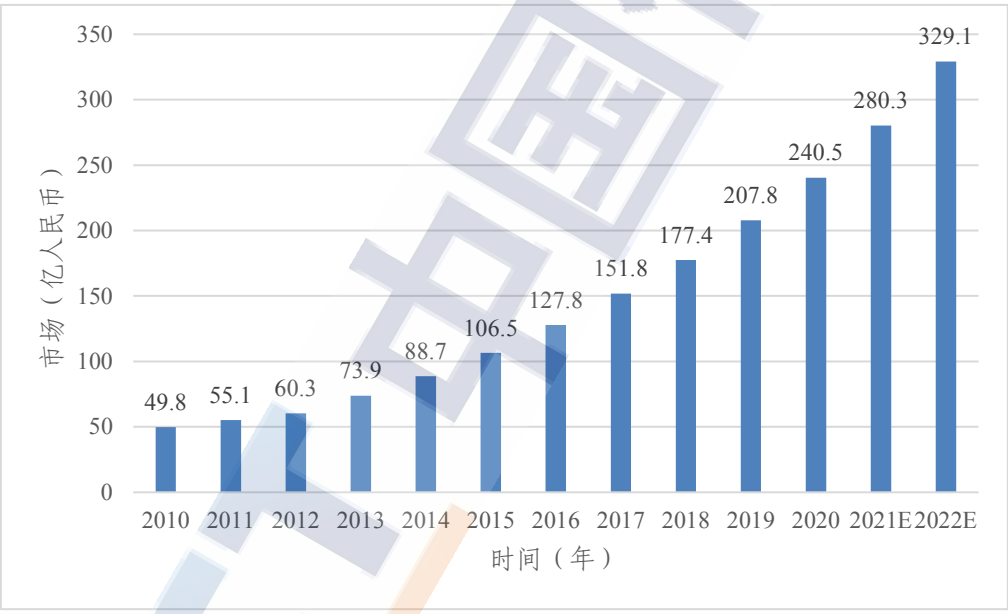
《银行业信息系统灾难恢复管理规范》（JR/T 0044-2008）第八章灾难备份系统建设时明确: 根据灾难恢复策略制定灾难备份系统技术方案,包含数据备份系统、备用数据处理系统和备用网络系统; 其次，为满足灾难恢复策略的要求，应对技术方案中关键技术应用的可行性进行验证测试，并记录和保存验证测试的结果; 最后，应制定灾难备份系统集成与测试计划并组织实施，通过技术和业务测试，确认灾难备份系统的功能和性能达到设计指标要求。

银行业信息安全监管还有包括但不限于:

- 《商业银行业数据中心监管指引》;
- 《网上银行业务连续性监控指引》;
- 《关于进一步加强银行业金融机构信息安全保障工作的指导意见》等。

金融领域确保业务连续性是重中之重，它关系到广大投资者和用户日常的投资行为和经济消费行为，任何由于数据库导致的非计划性停机，都可能引发巨大的经济损失和非经济性影响。

(三) 灾备市场与行业趋势



数据来源：2021中国灾备行业白皮书，英方研究院

图 1　2010-2022 中国灾备行业市场规模

智研咨询报告[4]显示，中国灾备行业市场规模从2010年的49.8亿人民币，增长至2018年近180亿人民币，预计至2022年中国灾备行业市场规模可达300亿以上。前瞻产业研究院的报告重点提到云灾备将成为未来趋势，云灾备市场规模从2013年的1亿元，快速增长到2018

[4]《2017-2022 年中国灾备行业深度研究及市场前景预测报告》

年的10亿元，预计到2022年规模达70亿元。Gartner预计到2021年，使用备份而非归档方式来管理企业长期的比例将由2017年的30%升至50%。国际灾备市场发展同样强劲，根据DataCore的2018年报告[5]显示，有20%的用户计划将存储预算的25%用于灾备方面。Marketsand Markets的相关数据也显示，全球备份和恢复市场总额将从2017年的71.3亿美元上升到2022年的115.9亿美元。

随着IT技术产品不断迭代，灾备应用场景从同机房本地备份容灾，向同城、异地及云端等更宏大的场景延伸；灾备技术从传统的存储复制技术，延伸到基于主机、操作系统、数据库、文件和网络等五大数据复制技术。灾备产品也正在不断拓展边界，涵盖传统系统备份、容灾和恢复；数据同步、分发、脱敏、副本管理；大数据管理与应用；数据库读写分离与容灾等。其中，数据库容灾架构在信息系统容灾架构中发挥至关重要的作用。

二、分布式数据库容灾架构

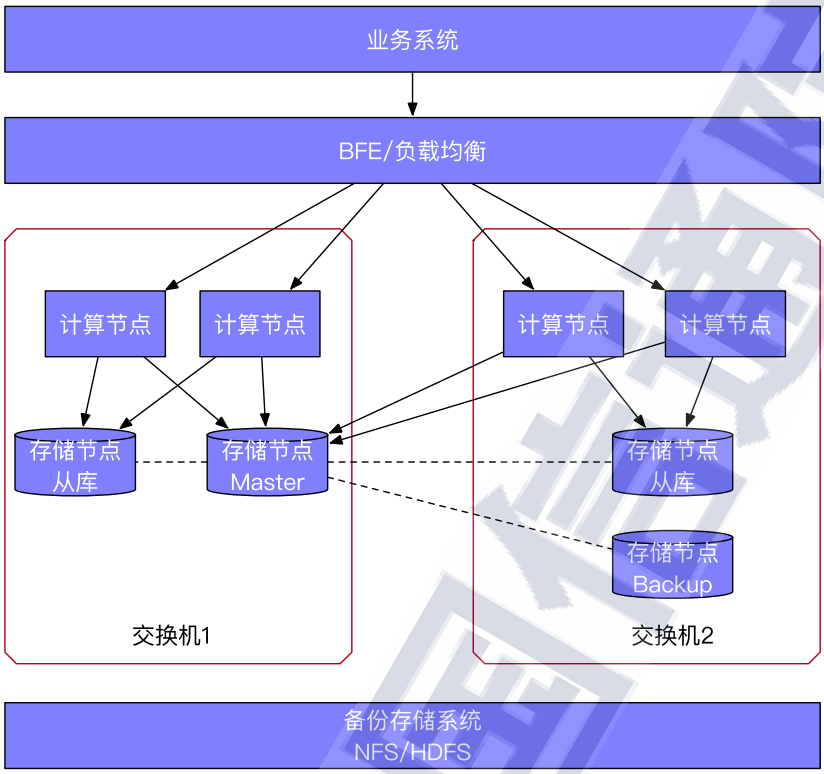
不同应用场景和业务需求下，分布式数据库的部署方式应该选择合理容灾架构，常见的分布式数据库容灾架构包括单中心容灾、同城互备、同城双活、两地三中心等。

(一) 单中心容灾

对于容灾要求不高的一些内部业务系统，可以只在单一的生产中心内部署。在这种模式下，数据库通过在该生产中心的多个不同可用区多实例部署，实现数据库服务高可用。不同可用区的数据库均能向

[5]《The State of Software-Defined Storage, Hyperconverged and Cloud Storage》

应用系统提供数据库访问服务。



来源：北京百度网讯科技有限公司

图 2 单中心容灾部署架构示意图

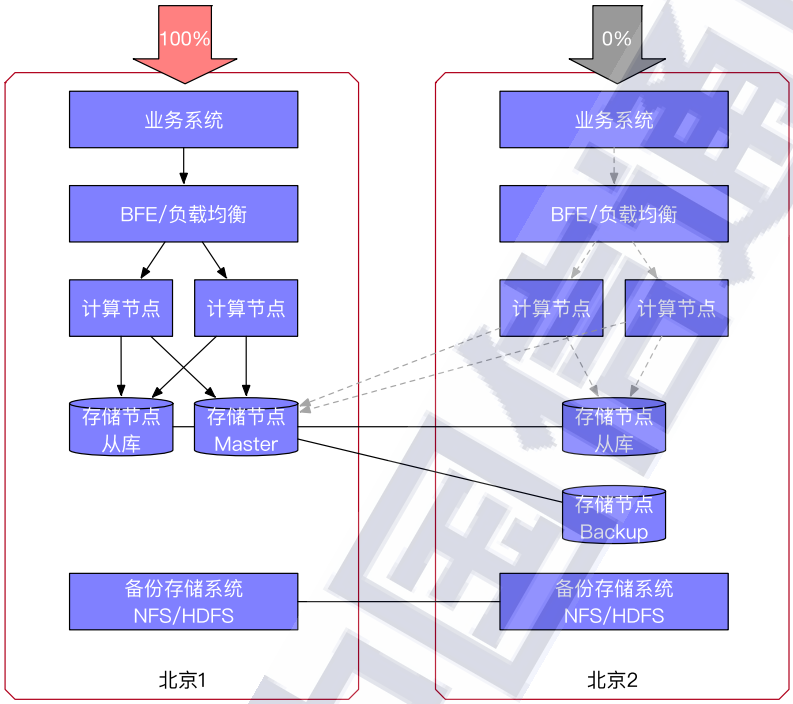
基于数据库的高可用机制，可以抵御部分节点甚至机房可用区整体故障，仍然保持数据库服务正常服务。但如果遇到数据中心级的灾难，则无法保持可用。

(二) 同城互备

同城互备模式需要在灾备中心内部署与生产中心业务系统相同资源配置，包括应用和数据库在内的完整业务系统。生产中心和灾备中心均能承载全流量业务压力。数据库系统在生产中心和灾备中心都需要冗余部署满足系统正常运行的全部组件，并配备满足全量数据存储和访问压力的存储和计算资源。正常情况下，只有生产中心投入运

行，灾备中心处于在线待机状态。当数据中心发生灾难时，灾备中心可以在短时间内切换并提供服务，快速实现业务止损。

异地互备模式部署架构图如下：



来源：北京百度网讯科技有限公司

图 3 异地互备容灾部署架构示意图

为了一定程度提高资源利用率，针对不同的核心业务系统，可交替设置主备中心。主备中心数据同步方式，可以采用强同步机制或异步同步机制。这取决于根据业务对数据一致性的要求，同时也受到数据中心距离带来的网络延时限制。

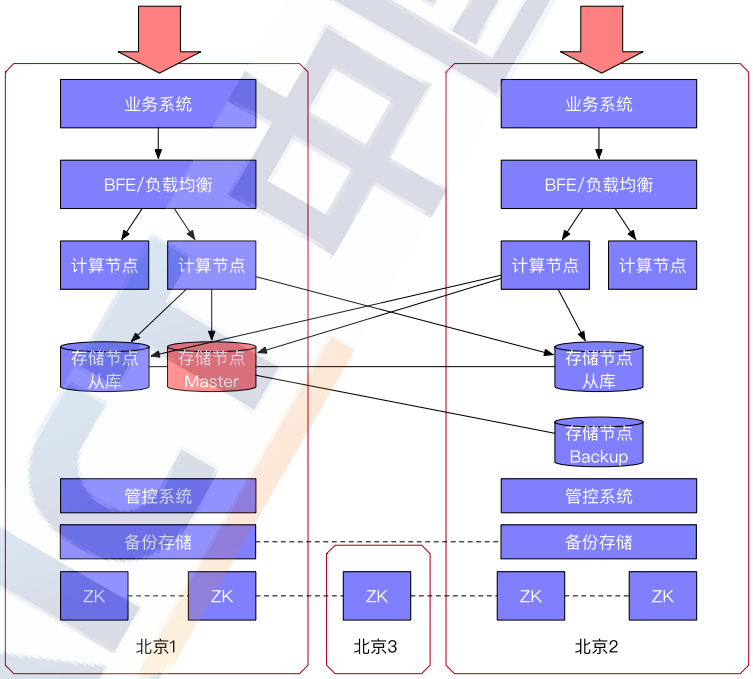
当数据中心和灾备中心在同城距离较近，网络延时较小，通常选择强同步，可称为同城互备；当数据中心间距离较远，跨地区进行传输的网络延时较大，可称为异地互备。

（三）同城双活

同城双活是双活技术与同城灾备中心模式结合的一种主流容灾架构。业务系统可以同时通过生产中心和灾备中心进行访问，无需指定特定的访问规则。数据库架构同时兼备异地互备模式的负载均衡和故障自动切换能力，且由于处于同城较近距离，两个数据中心的存储节点可以保持数据强一致。

当其中一个中心发生灾难时，通过接入前端的负载均衡调整，可将全流量输入对等的灾备中心；数据库同时自动进行切换，灾备中心的数据库集群承载全部查询请求。

同城双活的部署示意图如下：

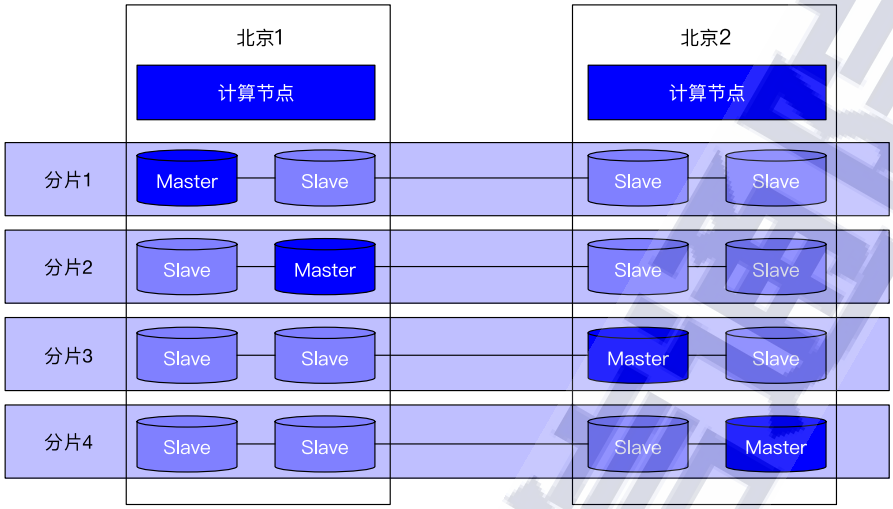


来源：北京百度网讯科技有限公司

图 4 同城双活容灾部署架构示意图

基于数据分布式架构可以对应用层提供透明的双活能力。以一个四分片的数据表为例，分片数据可以均匀分布在两个中心的数据库存

储节点中：



来源：北京百度网讯科技有限公司

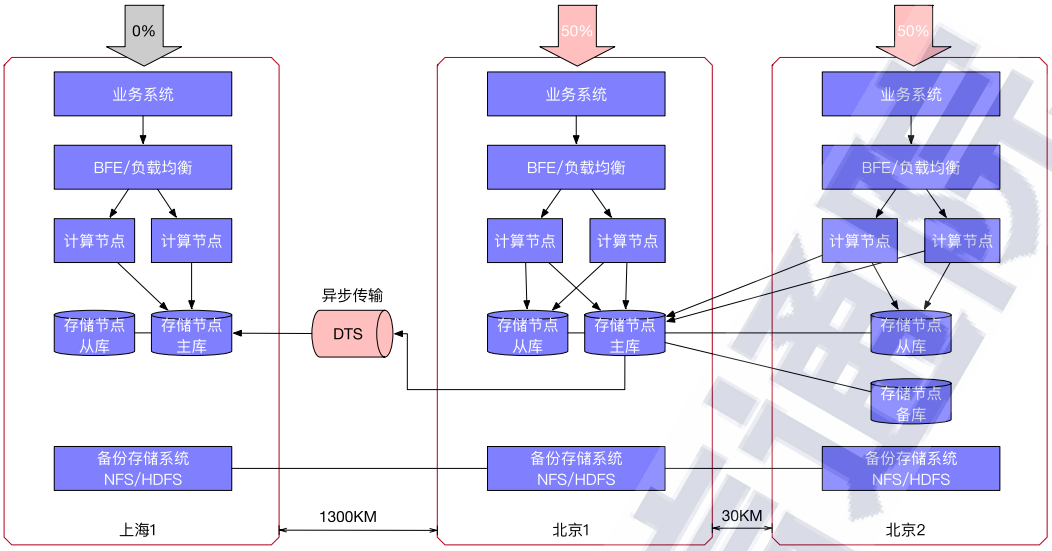
图 5 基于多个数据中心的数据分片示意图

在同城双活容灾架构下，数据库系统在生产中心和灾备中心也同样都需要冗余部署满足系统正常运行的全部组件，每一个计算中心都需要配备满足全量业务访问要求的存储和计算资源。

（四）两地三中心容灾

在同城双活或者同城互备的架构下，再增加一个远距离的容灾中心，可实现两地三中心的容灾架构。

该架构在同城容灾方案的基础上，获得了对地震、飓风等区域级灾难的抵御能力。由于异地灾备中心距离较远，所以数据同步一般考虑使用异步模式，可基于数据库异步同步功能实现，或者在应用层使用消息队列等组件进行业务数据异步同步，进而实现远距离异地机房的数据最终一致性。



来源：北京百度网讯科技有限公司

图 6 两地三中心容灾部署架构示意图

三、数据库容灾技术

数据库容灾技术与数据库的容灾架构紧密相关，在设计数据库容灾技术时，除了要考虑数据库容灾架构还要对数据的备份、恢复、传输等具体操作的实现细节。一套完整的数据库容灾技术既要有采用数据备份保护和恢复数据的功能，也要有保证数据安全、正确、高效传输的功能，同时还要有强大的应激防护能力。

(一) 数据备份与恢复

数据备份是以容灾为目的，为防止故障导致的数据丢失，而将全部或者部分数据复制到其他存储介质的过程。

备份过的数据往往通过压缩打包保存，提升存储效率。备份数据一般不能直接提供数据库服务，需要一套数据恢复操作流程进行备份逆操作，在现有数据库实例或基于空闲资源搭建新实例，覆盖原有数据，才能再次通过该数据库实例进行访问。所以数据备份也称为数据

冷备。

1.数据备份

数据库的备份技术从备份机制可分为物理备份和逻辑备份两种。物理备份是针对数据库的数据文件进行备份的方式。**物理备份往往是将数据库的数据文件连同备份时间窗口内的重做日志一同进行备份。**备份完成后，备份程序会重新执行重做日志，保持数据文件中数据的一致性。由于物理备份主要备份内容是复制数据库数据文件，能最大程度利用IO，备份效率很高。

逻辑备份是针对数据库的数据内容进行备份的方式。这种模式的优点是能灵活地选择备份范围，即可以选择指定的库表进行备份，还可以通过条件筛选备份内容。逻辑备份的模式下往往通过读取数据转换为“insert into table …”一类SQL语句，需要额外的解析和转换，同时还涉及大事务通过事务隔离性保障数据一致性，所以导致效率相对较低。

2.数据恢复

全量备份是基于时间点的数据库镜像，而增量备份是将此时间点扩展到持续的时间窗口。通过全量备份和增量备份，数据库可以恢复备份覆盖的时间范围内任意时间点的数据状态。同时数据恢复也可指定恢复目标，将恢复范围限制为部分库表，可大幅降低恢复时间开销。

(二)数据同步与传输

按照监管要求，核心业务系统生产中心的数据库数据能实时同步传输到异地的灾备中心。下面以MySQL数据库场景为例，介绍几种

不同的同步传输技术，供数据库容灾架构规划参考。

1.主从复制

MySQL主从复制[6]是一个异步的复制过程，主库发送更新事件到从库，从库读取更新记录，并执行更新记录，使得从库的内容与主库保持一致。

每一组主从复制的连接，都由三个独立的线程协作实现。在主库上为每一个连接到主库的从库创建一个二进制日志（以下简称“binlog”）输出线程。每一个发送给从库的SQL事件或者数据变更事件，都会基于binlog传输给从库。而每一个从库都会为同步创建一个I/O线程和一个SQL线程。I/O线程连接到主库并请求主库发送binlog事件到从库，读取到的binlog会更新到本地中继日志（以下简称“relay-log”）文件。而SQL线程读取到I/O线程写到relay-log的更新事件后，在数据库中进行执行，从而完成数据从主库到从库的同步。

2.半同步复制

半同步复制是MySQL 5.5版本引入的机制[7]。半同步复制出现前，虽然异步复制可以满足主从实例之间的数据同步要求，但如果主库崩溃，将从库提升为主库时，原来的主库上可能有一部分已经提交的数据还没有来得及被从库接收，主从不一致将导致切换后的新主库丢失这一部分数据。

半同步复制改进了事务提交的逻辑。客户端在主库上写入一个事

[6]MySQL 官方手册：《Replication》

[7]MySQL 官方手册：《Semisynchronous Replication》

务时，需要等待从库接收到主库相关的binlog数据，且主库接收到从库的应答之后，客户端才能收到事务成功提交的消息。

早期的半同步复制存在一些缺陷。主库在等待应答的过程中，存储引擎内部已经提交的事务，只是阻塞返回客户端消息而已，此时如果有其他会话对该会话事务修改数据进行查询会查询到数据。如果此时出现主库故障转移，非发起数据库提交的客户端可能会出现幻读。所以MySQL 5.7版本对半同步进行了优化，称为增强半同步复制。优化后一个事务在存储引擎内部提交之前，必须要先收到从库的应答确认，否则不进行事务的最后提交，从而解决上述提到的幻读问题。

3.组复制

MySQL在5.7版本正式推出组复制（MySQL Group Replication[8]，以下简称“MGR”）机制。MGR集群由若干个节点共同组成一个复制组，一个事务的提交，必须经过组内大多数节点决议并通过才能得以提交。

引入组复制，主要是为了解决异步主从复制和半同步复制可能产生数据不一致的问题。组复制依靠分布式一致性协议实现了分布式架构下数据的最终一致性，提供了MySQL集群的多主写入方案。

MGR技术与Oracle RAC类似，对集群网络的要求非常高。网络延时和不稳定会严重影响MGR集群的正常工作，所以多用在单数据中心的内网环境中，对于主备和多活容灾架构下异地同步的场景，存在一定的短板。

[8]MySQL 官方手册：《Group Replication》

4.分组强同步

半同步复制机制保障了主库故障切换时事务数据能够在至少一个从库中持久化存储，保证切换过程不丢失最新数据。随着数据库集群规模逐渐增大，同城和异地多机房灾备架构对同步的要求也愈发提高。当跨多机房部署的集群出现大规模故障，例如机房故障或专线故障时，主库和完成接收binlog数据的从库节点可能同时出现故障。因此在半同步的基础上，出现了分组强同步机制。

分组强同步机制能够保证在跨机房的场景下仍然保持高可用和强同步的能力。任何一个集群的从库可以分成若干组，在每一组中，只要有一个从库返回成功，则认为该组复制成功。当所有的组都复制成功，主库的事务才能完成提交。

分组强同步复制算法可以保证已经提交成功事务的数据不丢失，修复了MySQL原生半同步可能丢失数据的隐患，确保在主库发生故障情况下，不会因为二进制日志丢失导致从库丢失数据，进一步提升了数据的可靠性。

5.云数据库数据同步服务

为了与数据库产品配套，云平台供应商和数据库厂商推出数据传输服务（Data Transmission Service，以下简称“DTS”），该服务用于在异构数据库之间进行数据迁移、数据同步和数据订阅。DTS 支持在业务不影响源数据库服务的前提下进行数据库迁移，利用实时同步通道构建异地容灾的高可用数据库架构。DTS 往往支持在主流数据库之间进行结构迁移、全量数据迁移和实时增量数据同步，其迁移同步任

务还可按照同步范围并行进行同步。

数据传输服务在异地灾备场景下也被作为异步同步的重要方案。

来源：北京百度网讯科技有限公司

图 7 数据传输服务架构示意图

（三）故障自动切换

当分布式数据库各类节点出现故障时，其监控系统应该能实时感知到故障种类和范围，包括各类节点的进程故障、服务器故障、磁盘故障和网络故障等，都可以依据预案配置，自动进行故障切换。

主备容灾架构下，容灾机房内会建设一套与生产机房相同规模的服务。如果生产中心出现灾难而不可用，数据库管理系统应该能自动将数据库服务切换至灾备中心。在异地容灾架构下，数据库甚至能够抵御地理级灾难，如地震洪水等。该类灾难可能会影响整个城市区域，使得同城机房均不可用，从而将服务切换至异地容灾中心。

1.集中式架构

以SQL Server数据库为例，介绍集中式架构数据库的典型故障切

换技术。SQL Server采用SQL Server Always On可用性组来支持对一组分散的数据库实施故障转移。一个可用性组支持一组读写主数据库，以及一至八组对应的辅助数据库(包括一个主副本和最多四个同步提交辅助副本)，每个副本承载一组辅助数据库，同时也是可用性组的潜在故障转移目标。发生故障转移时，数据库通过一组独立的服务器故障转移群集服务，将实例的资源所有权转移到指定的故障转移节点。然后SQL Server实例在故障转移节点上重新启动，使数据库恢复如常。无论在任何时刻，故障转移群集中只有一个节点可以承载故障转移群集实例和基础资源[9]。

Always On可用性组主副本将每个主数据库的事务日志记录发送到每个辅助数据库，每个次要副本则缓存事务日志记录，然后将日志记录应用到相应的辅助数据库。主数据库与每个连接的辅助数据库独立进行数据同步。因此一个辅助数据库可以挂起或失败，但不会影响其他辅助数据库，一个主数据库也可以挂起或失败，也不会影响其他主数据库[10]。

2.分布式架构

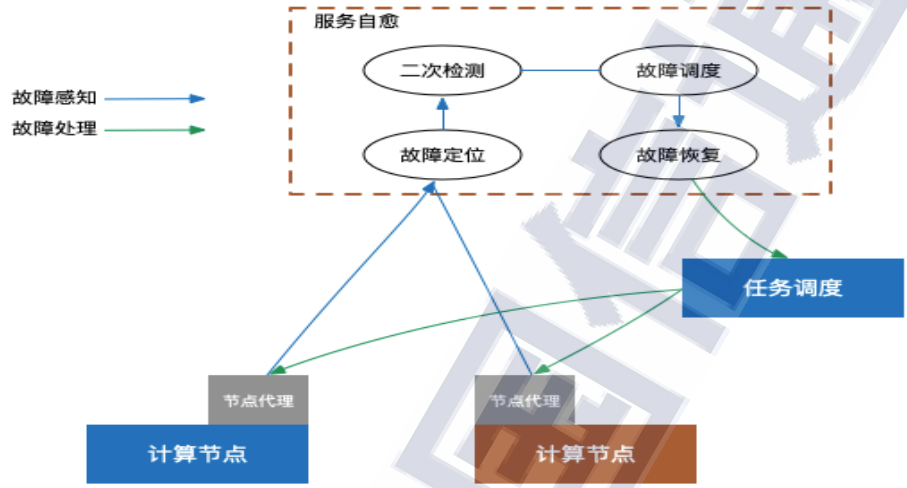
分布式数据库架构由于进程、磁盘和服务器等故障，往往导致集群的少量节点实例不可用，应对措施主要是通过冗余和副本节点进行替换止损，替换过程中可能出现主备切换或主从切换；对于交换机、路由器等网络设备发生故障，除了导致部分节点实例不可用外，还可能出现集群分裂情况，因此分布式数据库需要建立应付各种类型和规

[9]SQL Server 官方文档《Windows Server 故障转移群集与 SQL Server》
[10]SQL Server 官方文档《什么是 Always On 可用性组？》

模故障的容灾能力。

（1）计算节点故障切换

当其中一个计算节点出现故障，流量以秒级切换至其他计算节点。整个切换过程对用户透明，应用代码无需变更，应用进程无需重启。



来源：北京百度网讯科技有限公司

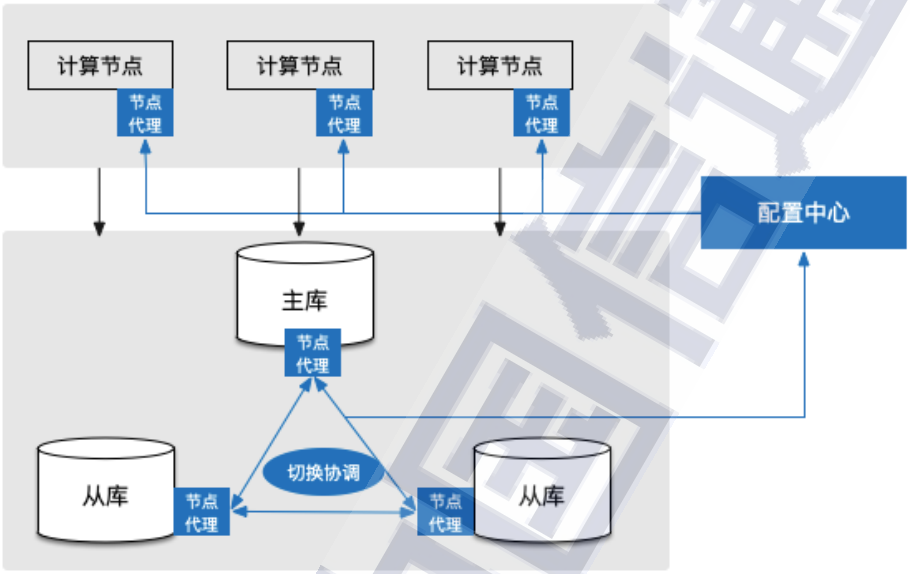
图 8 计算节点故障切换示意图

计算节点自愈分为故障感知和故障处理两部分。故障感知是指通过节点代理的定时任务定期执行自愈监控项采集任务，对计算节点的监控指标进行采集，并上报至服务自愈模块，该模块对节点的监控数据进行分析，对可能的故障信息进行定位和二次检测，若确定为故障则发起故障处理任务。故障处理是指服务自愈模块检测到故障节点，任务调度模块创建自愈任务，自动对故障节点进行恢复处理，处理完成后故障节点重新上线恢复正常服务。

（2）存储节点故障切换

分布式数据库采用多副本保存数据，存储节点通过多副本方式构

成集群。最常见的集群模式是主从集群，即一个主节点负责写入，并基于一定的一致性算法同步至其他从库。当对外提供数据库服务时，主库承担读写服务，从库提供只读服务。当主库节点故障时，系统会自动发现并尝试恢复主库，如果主库无法恢复则发起主从切换。



来源：北京百度网讯科技有限公司

图 9 存储节点故障切换示意图

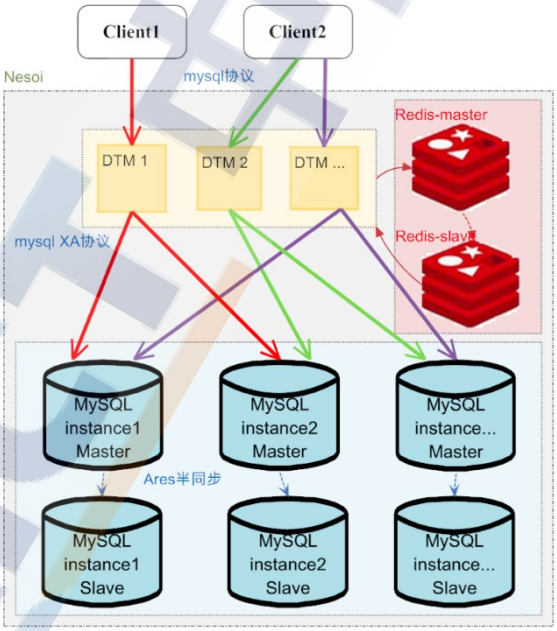
切换协调模块为切换核心模块，负责存储节点健康诊断、切换仲裁与协调，并变更集群的拓扑信息。如上图所示，数据库实例的三个节点代理构成了切换的协调者。节点代理通过与决策集群通信获取其托管的集群元信息，并借助集群取得集群中其它节点代理的通讯方式。

(四)分布式事务容灾

金融类业务数据模型复杂度较高，往往需要多维度进行数据处理和分析。进行业务系统分布式改造时，分片策略往往无法避免一个事务可能跨越多个数据分片的情况，需要通过分布式事务保证数据的强

一致性。为保障分布式事务在跨节点处理时事务的原子性和一致性，一般使用分布式协议处理。常用两阶段提交、三阶段提交协议保障事务的原子性；使用Paxos、Raft等协议同步数据库的事务日志从而保障事务的一致性。支持分布式事务是金融级分布式数据库产品的核心特性，分布式事务的容灾能力是分布式系统容灾能力的重要考量指标。

百度分布式数据库GaiaDB-X采用基于优化的XA协议及两阶段提交算法实现分布式事务机制。其优势有：1）自研DMVCC[11]算法，解决分布式全局读一致性问题；2）解决MySQL原生XA在事务一致性和持久性上的缺陷；3）在宕机等故障场景下，能够基于持久化的全局事务状态对悬挂事务进行提交或回滚，保证分布式事务的容灾恢复能力；4）支持备份恢复的事务一致性、死锁检测等高级功能。



来源：北京百度网讯科技有限公司

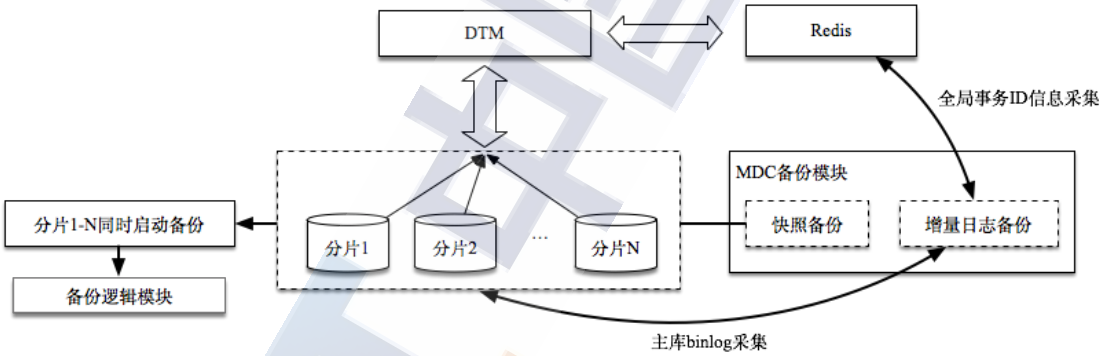
图 10　GaiaDB-X 分布式事务机制示意图

[11]*DMVCC：Shao J, Yin B, Chen B, et al. Read Consistency in Distributed Database Based on DMVCC[C] IEEE, International Conference on High PERFORMANCE Computing. IEEE, 2017:142-151.

上图为GaiaDB-X的分布式事务架构，其中分布式事务处理器（Distributed Transaction Manager，以下简称“DTM”）负责SQL路由并管理分布式事务，是业务端访问数据库服务的入口。DTM使用Redis存储多版本的读视图基线并存储数据节点的分布式事务状态信息。

在宕机等故障场景下，全局事务状态持久化在高可用Redis集群中，故障自愈后，新节点能通过Redis恢复悬挂的事务，决定事务应该提交或回滚，保障分布式事务的高可用能力。

当分布式表进行备份和恢复时，如何保证恢复数据的一致性也是考验分布式数据库的难题。GaiaDB-X实现了基于全局事务点的备份恢复机制。



来源：北京百度网讯科技有限公司

图 11 基于全局事务点的备份恢复机制示意图

所有分片同时通过备库上定时任务启动备份任务独立进行备份，根据备份结果获取全局事务标识（Global Transaction ID，以下简称“GTID”），并从主库日志中解析得到备份快照点及对应分片的全局事务信息，与备份数据一同备份。这样基于快照点与全局事务信息的对应关系，即可在备份恢复时保证各分片的全局事务一致性。

（五）应用应激防护

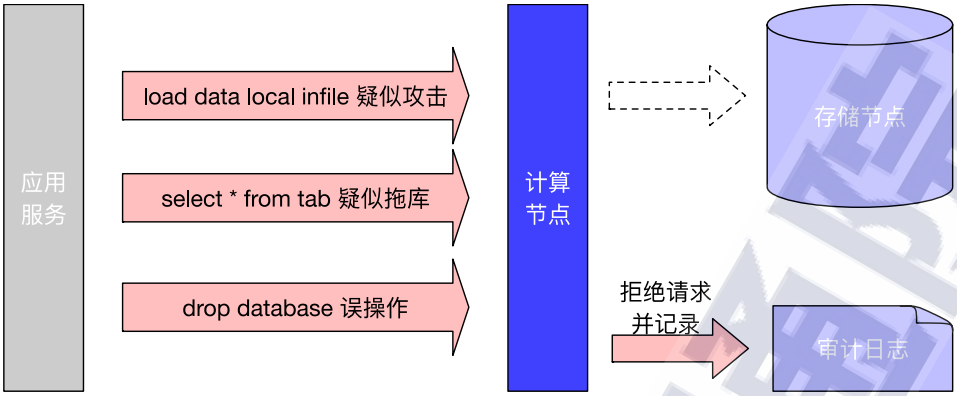
1.过载保护

为了保障数据库服务不受业务系统突发过量压力影响，导致系统稳定性下降，数据库需要提供对请求流量进行控制的功能，保障数据库服务在极端应用负载情况下仍能稳定运行。当监测到连接可用性下降时，会触发进入限流模式，基于实时查询量对主库和从库的访问量进行限制。限制策略有连接数限制、查询量限制和处理时间限制等。

2.SQL 入侵防御

SQL入侵防御是对用户发送的SQL进行语法解析，过滤非法SQL的一种安全能力。数据库产品一般通过中间件或基于计算节点集成防火墙功能，对非法SQL进行过滤、阻断和报警，有效的预防SQL注入或恶意非法攻击。对于不同SQL，防火墙采用拦截模式或告警模式进行处理: 拦截模式是指当发现非法SQL时，防火墙会自动阻断其执行，并发送报警通知；告警模式是指当发现非法SQL时，防火墙会允许其继续执行，但会发送告警通知。

基于SQL入侵防御技术，数据库可以有效识别和拦截外部的SQL攻击行为，并记录日志，供事后追查。例如各类SQL注入攻击模式，包括脱库、高危SQL等，SQL防火墙都能实现有效拦截。



来源：北京百度网讯科技有限公司

图 12　SQL 防火墙示意图

3.数据回滚

在存储限额范围内，数据库将删除的数据库表移动到数据回收站保存。对于误操作或者需要恢复删除的情况，可以对数据库表进行DROP操作后进行快速“闪回”。数据回收站的数据会自动按照清理策略进行清除，不会影响正常数据库服务。

4.弹性扩容

当分布式数据库的负载不断增加，或出现短期的业务浪涌，可以使用弹性扩容功能对计算节点和存储节点进行扩容，保障数据库服务平稳运行。

（1）计算节点水平扩容

当集群的处理能力不足的时候，分布式数据库支持在线增加计算节点，扩展服务能力。同时，当集群的计算节点资源利用率较低时，支持降低集群规模，减少计算节点数量，降低服务能力，提供服务能力弹性扩展能力。

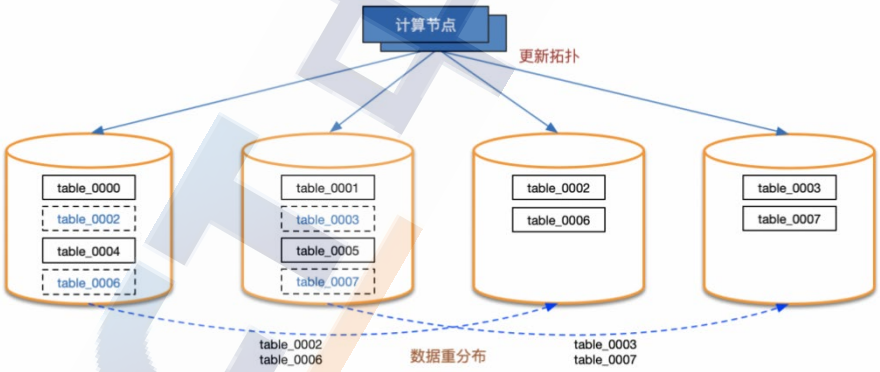
由于在分布式数据库系统中计算节点往往被设计为无状态，通过

新增或减少计算节点数量，并在配置节点中进行相关元数据的更新即可完成计算节点的调整。

（2）存储节点水平扩容

存储节点水平扩展通过增加存储节点线性提升集群整体容量和吞吐性能，分布式数据库的控制台可以发起扩容操作。以百度分布式数据库为例，该操作主要分成四个步骤：1）创建新的存储节点；2）重新分布数据，并保持新节点和原节点数据实时同步；3）变更计算节点拓扑，切换流量到新节点；4）清理原节点上的数据。

在整体扩容过程中，第三步会对业务有秒级的连接闪断（一般应用层具有重连机制），因此分布式数据库往往提供手动触发第三步切换操作，运维人员可选择业务低峰期完成切换。



来源：北京百度网讯科技有限公司

图 13 存储节点水平扩展示意图

四、分布式数据库容灾方案

分布式数据库容灾设计核心思想是就是充分利用分布式多副本数据一致性协议原理，并结合各种意外情况的具体特点，找到对应的

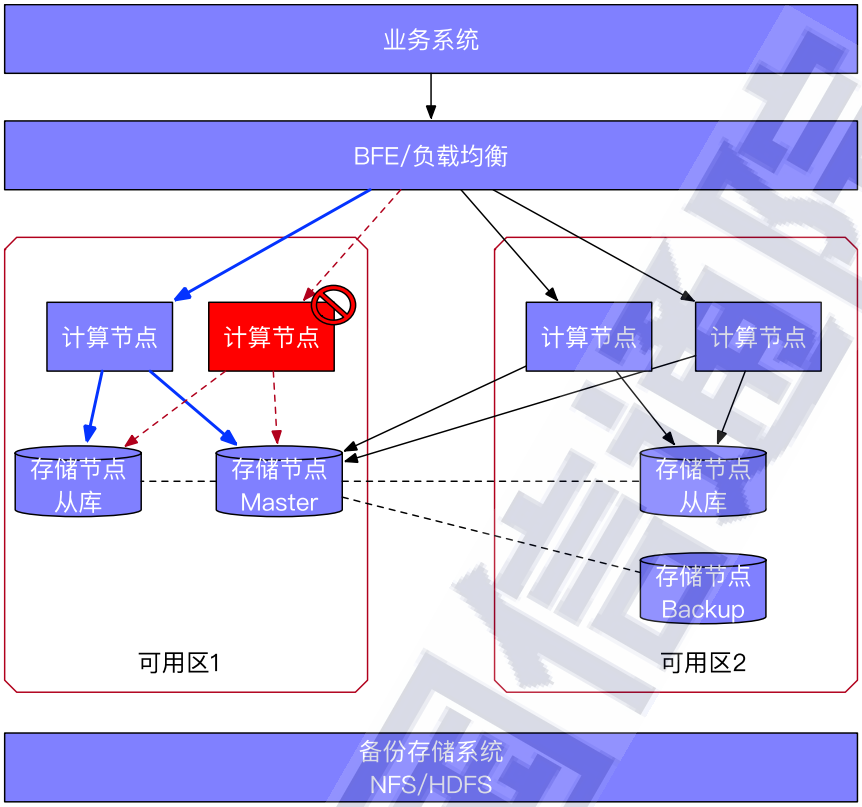
解决方案。及时发现故障，准确分析故障原因在分布式数据库容灾过程中的关键。集群中任何节点的软硬件故障，或者任何节点之间的网络连接故障，都可能影响集群的正常工作。由于其天然的复杂性，这个环境发生故障的几率并不低，因此分布式数据库系统需要将节点软硬件故障和网络故障当作常见情况来处理，而不是当作极低概率的事件来处理。下文从节点故障、网络故障、数据中心故障等维度重点介绍数据库容灾应对方案。

(一) 节点故障

为了能描述系统容灾的过程，下面先对分布式数据库的常见节点故障处理过程进行介绍。

1.计算节点故障

分布式数据库的计算节点多采用多实例部署，且自身不维持状态。所以当一个计算节点故障后，负载均衡通过健康检查识别故障，自动把请求分发到其他的计算节点上。



来源：北京百度网讯科技有限公司

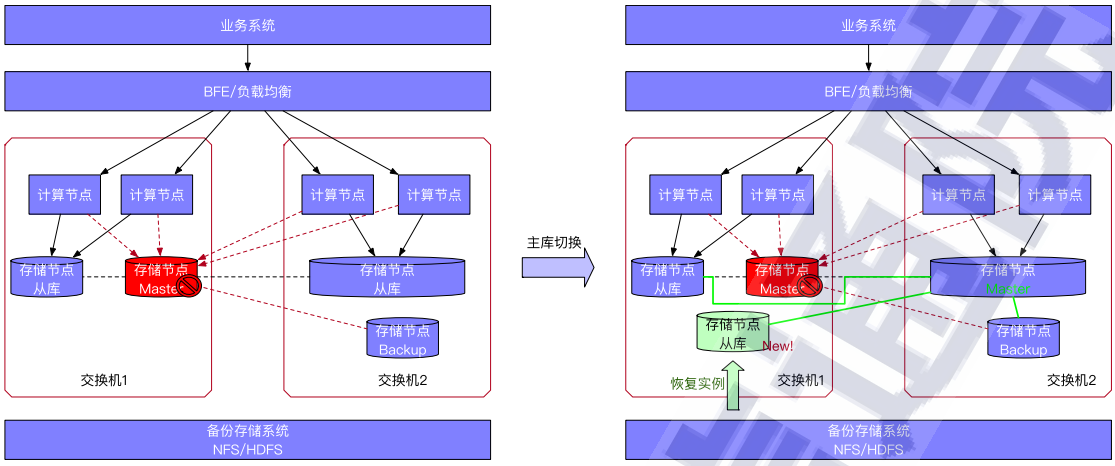
图 14 计算节点故障转移切换示意图

待该节点自愈恢复后，负载均衡检测到节点正常服务，会自动将请求重新均衡到该计算节点上。

2.存储节点主库故障

当分布式数据库的存储节点集群出现主库故障，集群的健康检查模块会监测到主库故障，并通过多次探测确认故障。如果主库故障确认，系统会控制存储节点集群进行主从切换，将一个从库节点选举为新的主库节点，并调整集群的主从同步关系。同时集群拓扑的元数据会进行更新，推送到集群的各节点，包括查询路由相关信息。这样应用层的写操作会发送到新的主库。此时由于存在一个故障的副本节点，根据集群的副本策略和恢复机制，恢复系统会触发一个恢复任务，重

新恢复一个新的从库节点加入集群。

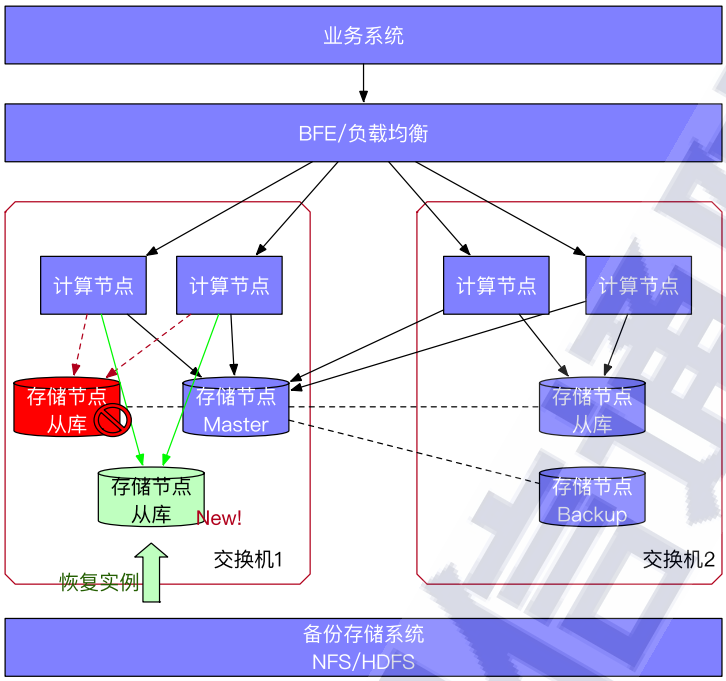


来源：北京百度网讯科技有限公司

图 15 主库节点故障转移切换示意图

3.存储节点从库故障

当分布式集群的一组主从出现从库节点故障，集群的健康检查模块会监测到从库故障，并通过多次探测来确认故障。如果确认从库故障，系统会将该从库进行下线处理。同时集群拓扑的元数据会进行更新，将该故障的从库节点从集群信息中删除，并推送到集群的各节点进行更新。查询路由相关的信息更新后，应用层的读操作不再发送到故障的从库。此时由于存在一个故障的副本节点，根据集群的副本策略和恢复机制，系统的恢复系统会触发一个恢复任务，重新恢复一个新的从库节点加入集群。



来源：北京百度网讯科技有限公司

图 16 从库节点故障转移切换示意图

(二) 网络故障

1.对查询的影响

专线网络抖动异常一般会带来延时和丢包。GaiaDB-X集群的计算服务节点、存储服务节点、管控节点之间的通讯都有完备的超时重试机制，少量的延时丢包不会影响查询性能，跨专线的主从延时会增大，待网络质量恢复后即恢复正常。

2.对写入的影响

采用分组半同步的集群，网络延时和丢包可能会造成同步延时甚至同步中断重连，影响写操作提交，导致部分事务提交失败。一般来说，分组半同步有两种应对策略:

一是支持配置同步退化超时时间，即延时超过一定时间后强同步退化为异步同步，可以减少链路异常对应用写入操作的影响，但同步退化后降级为异步同步，主从数据的强一致性已经不能保障。此时如果网络抖动蔓延成网络中断故障，出现存储节点集群切换后会出现数据丢失，即RPO>0。

二是应用需要最大程度保障数据一致性，由于网络抖动情况复杂，所以采用牺牲这段异常时间段的写入功能可用性，保持强同步策略。所有写入全部会被网络异常阻塞，等待人工处理或者网络恢复；实际情况应该根据数据库的业务要求采用不同的策略配置。

（三）同城双活生产中心灾难

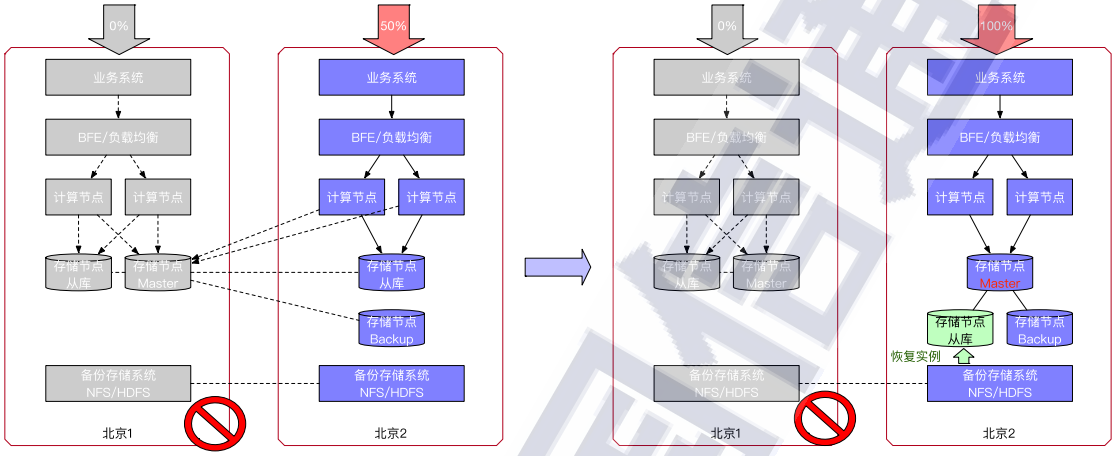
以同城双活容灾架构为例，此处介绍遇到机房级灾难情况下系统的应对措施。

当生产中心机房出现严重灾难，导致大面积服务器和网络故障。这时灾备中心的数据库集群会监测到生产中心的计算节点和存储节点不可用，开始进行容灾应急处理。

首先系统会对生产中心的计算节点进行下线处理。由于生产中心的计算节点已经全部不可访问，系统会在元数据中将所有异常节点进行剔除，并调整负载均衡配置，生产中心的所有查询流量都只会发送到容灾中心的计算节点。同时系统对生产中心的存储节点进行下线处理。对于主库在生产中心的集群，按照存储节点主库故障的处理流程进行自动主从切换，将灾备中心的从库切换为主库；对于主库在灾备中心的集群，则按照存储节点从库故障的处理流程进行节点下线。调

整完成后更新集群元数据，所有存活的计算节点都会调整路由，将读写操作发送到新的容灾机房的存储节点集群；最后根据容灾策略，系统会对复本率不足的主从集群补充从库节点。

整个容灾切换的处理示意图如下：

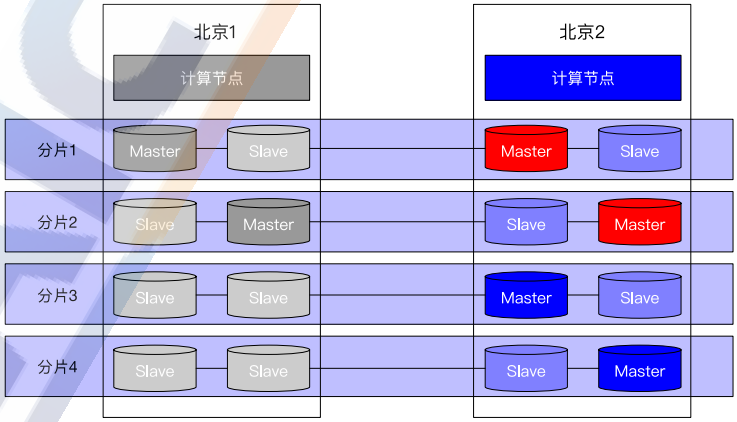


来源：北京百度网讯科技有限公司

图 17 同城双活生产中心故障转移切换示意图

此时容灾中心的负载均衡入口可以承载全流量的业务压力，业务系统可自动或手动将流量全部转移到灾备中心的数据库集群。

以一个四分片的分布式数据库为例，主从切换后，分片拓扑调整如下（不包括补充副本）：



来源：北京百度网讯科技有限公司

图 18 主从切换后分片拓扑调整示意图

在同城双活的容灾架构下，对分布式数据库容灾指标的特性进行分析。

1.数据库服务 RTO

数据库服务的主要节点是计算节点和存储节点。下面针对两种节点的故障恢复进行分析。

计算服务节点一般采用多实例部署，节点本身不保存数据和状态。通过负载均衡设备进行流量均衡；当一个计算节点发生故障后，往往可以在秒级内完成节点的故障转移。落到该故障节点的应用访问一般重试后即可恢复。

存储服务节点一般采用多分片和多副本的模式组建集群。出现存储节点故障情况下，数据库服务完成恢复时间包括四部分:

- 集群故障监测时间

分布式数据库采用多重监测机制来保障故障感知的准确性:

节点监测：存储节点实例的可用性监测（服务进程状态、内存使用情况、磁盘设备状态）。单个存储节点的心跳是定时检测，当连续检测到一定次数的异常心跳，即判断节点故障。建议心跳连续验证次数配置不能太低，否则容易出现误切换；另外对于异地机房节点，建议增加连续验证次数，以防线路抖动触发误切换；

同步监测：获取集群各存储节点的同步关系的信息数据。通过分析所有存储节点的主从关系，验证其组成的拓扑关系正确性。这种监控不需要连续检测，一次检测即可发现集群异常；

集群监测：监测存储节点的存活和同步状态，保障切换的最小集

群规模并避免反复切换。建议连续多次监测失败才能确认异常。

通过多种监测机制完成确认集群异常，建议通过配置测试各阶段的合理重复检测次数，使整体检测时间保持在10至20秒合理范围，过短的监测时间容易出现误感知导致误切换。

- 主从切换准备

当存储服务主库节点异常时，需要启动主从切换流程进行故障转移。主从切换前，往往需要在仍然存活的副本中选举一个新主库，形成新集群拓扑。在新集群进行切换前，各存储节点根据同步机制不同，可能存在数据不一致的情况。所以接下来需要等待所有节点完成数据交换补齐，将各节点数据同步到最新状态。

在一般情况下，所有节点不存在较高同步延时，数据补齐准备工作会在1至2秒内完成，但对于负载很高的集群，主从之间会存在一定的同步延时，则切换准备时间会在几秒到几十分钟不等。造成较高同步延时的原因有:

存在批量大事务执行，例如批量导入和删除任务，建议将任务放在业务低峰期或拆解为多批次小数据量执行。

有突较高业务压力，建议对重点数据表进行分表拆分。

- 主从切换

节点全部完成同步后，所有存储节点的数据一致，此时重新调整主从关系，切换一般秒级即可完成。

- 集群路由调整

完成切换后，配置节点中的集群拓扑结构会进行相应的调整，并

同步到所有计算节点，时间与元数据同步周期有关，一般在数秒左右。

正常情况下，基于分布式数据库的多副本同步复制机制，在同步延时较小的集群上，RTO故障恢复时间应该可以保持在30秒到1分钟左右。

2.数据库服务 RPO

基于强同步机制的数据库服务可以实现数据无丢失，即RPO=0。实际的数据库应用场景中，RPO的能力取决于采用何种同步机制策略。如果应用系统容忍丢失少量数据，则可以采用异步同步机制，这样还可以提高集群写入性能，但RTO>0；如果应用系统严格要求数据完整性，则可以采用强同步机制，虽然集群写入性能会受到一定影响，但能保证故障切换后不丢失数据，即RPO=0。一般来说，使用强同步后，事务提交的延时会增加大概一个集群节点网络间往返时延（Round-Trip Time，以下简称“RTT”）的最大值。以同城双活容灾架构为例，最大的RTT是两个同城机房间的网络交互延时，一般约为1至2毫秒。即开启强同步后，每个事务都会增加1至2毫秒的时间开销。

3.专线延时对查询的影响

分布式数据库的同城双活容灾架构下，数据是分布在两个机房的。如果查询请求所需的存储服务节点不在当前机房，就需要一次跨专线的访问。与跨机房的强同步类似，一次交互也需要额外的1至2毫秒时间开销。

减少跨机房查询的有效方式是业务系统设计时就考虑分布式数据库适配，采用单元化和服务化设计，可以将跨机房的查询情况降低

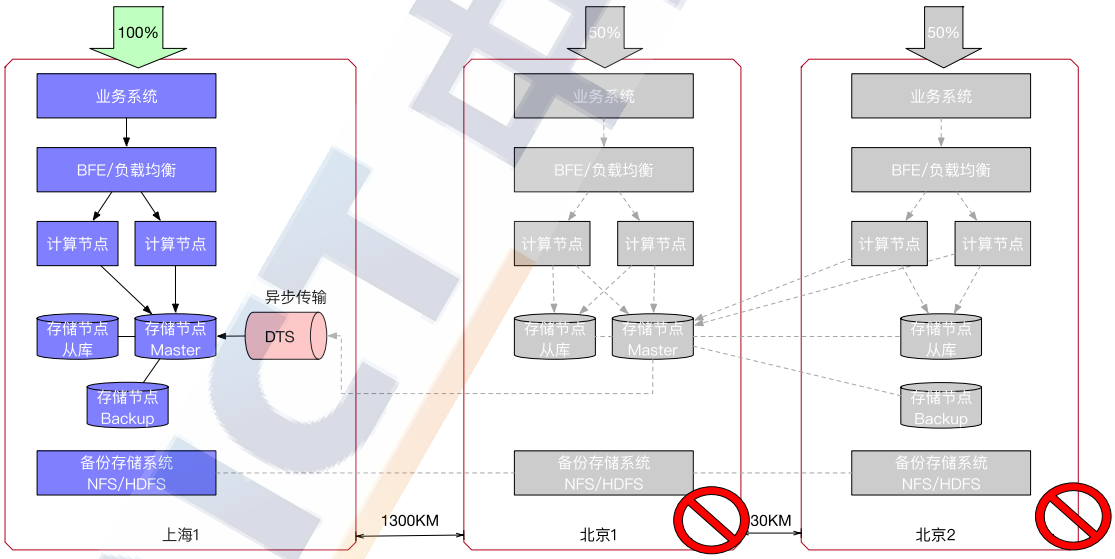
90%以上。如果业务系统没有进行单元化改造，或者业务系统是单体架构，采用数据分布式方式构建系统，极端情况下跨库访问的概率可能会超过50%。

（四）两地三中心主区域灾难

以两地三中心容灾架构为例，介绍遇到地理级灾难情况下系统的应对措施。

当出现城市级重大灾难，位于该城市的生产中心和灾备中心出现大面积的服务器和网络故障，所有数据库节点均受到影响。异地灾备中心的数据库集群会监测到生产中心和同城灾备中心的计算节点和存储节点不可用，开始进行容灾应急处理。

完成切换后的系统架构如下：



来源：北京百度网讯科技有限公司

图 19 两地三中心容灾切换示意图

此时异地容灾中心已经可以对外提供服务，可将应用流量映入异地容灾中心，恢复服务。

五、总结与展望

数据库技术历经六十余年发展，新架构层出不穷，技术不断完善。本文首先介绍数据库容灾和备份技术概念，然后结合集中式和分布式架构对多种数据库产品的容灾备份技术进行对比分析。从系统架构来看，集中式数据库架构在解决方案和产品能力上表现更为成熟，但在技术上缺乏突破，略显保守；分布式关系型数据库虽然出现时间较短，但创新的技术架构和扩展能力使其在市场中更具活力与挑战。

从数据库系统的容灾架构看，金融业务系统开始从单中心容灾架构，逐步向同城互备、同城双活或异地多活等容灾架构进行演化。通过分析分布式数据库产品在不同容灾架构下的容灾备份功能特性，可以看到在容灾性能指标、数据可靠保障、应用场景支持等方面，分布式数据库都已经有较为完善的容灾解决方案，技术发展潜力巨大。

随着分布式数据库技术不断发展，容灾技术应向着智能化、云原生为主的方向发展。

(一) 混合业务负载降低容灾复杂度

混合业务负载数据库，可以同时支撑联机事务处理和联机分析处理的业务场景，打破了事务处理和数据分析之间的隔阂，同时避免了在传统架构中，在线数据库与离线数据库之间大量的数据交互。实现一份数据，面向多种场景的统一数据库服务架构。这种模式下对数据流的简化和存储架构的统一，也降低了数据库系统容灾方案的复杂度。

(二) 人工智能改善容灾处置灵活性

随着分布式数据库架构不断演进，数据库组件功能进一步细化，

数据库系统将会形成完整的生态体系，节点的管控由目前的基于规则模式逐步演化为基于人工智能算法的协同模式。数据库系统除了能自动应对基础设施故障外，还能通过对各节点流量进行大数据分析和预测，实现系统级和应用级的异常检测及灾难的预警和隔离，同时在灾难发生时提供更细粒度的处置能力。

（三）云原生实现容灾过程可编排

云计算拐点已至，云原生成为驱动业务增长的重要引擎。云原生数据库基于抽象的云原生基础架构，提供Serverless服务，以支持应用为核心，而传统基础架构下的数据库产品往往是以存储为核心。

在容灾场景里，传统架构数据库的容灾过程参与者主要是主机、存储、网络和软件；而在云原生数据库平台，容灾参与的对象是云资源和在多云之间互相流转的数据，利用云原生编排能力，结合业务场景有望实现容灾高度自动化。

（四）混沌工程提升容灾架构韧性

现实世界中的各类灾难往往是无法避免的，目前的故障演练一般通过断网断电来测试系统的恢复能力，而分布式系统的复杂度需要一种更灵活更具有探索性的方案验证系统可用性的边界。

“混沌工程”是一种可实验的、基于系统的方式，可以用来分析分布式数据库系统中的混乱问题。通过不断实验，观察分布式数据库系统的行为和反应，了解系统实际能承受的韧性边界。分布式数据库架构下，节点间的交互和依赖极端复杂，系统的容错容灾技术特性各

异，传统技术已无法有效完成容灾演练测试，引入混沌工程可以更好的验证并提升系统容灾架构的韧性。

参考文献

[1] 2021 中国灾备行业白皮书. 国际灾难恢复（中国）协会（DRI China）.

[2] 2017-2022 年中国灾备行业深度研究及市场前景预测报告. 智研咨询集团.

[3]银行业信息系统灾难恢复管理规范. JR/T 0044-2008.

[4]证券期货业数据分类分级指引. JR/T 0158-2018.

[5]证券期货经营机构信息系统备份能力标准. JR/T 0059-2010.

[6]分布式数据库技术金融应用规范 灾难恢复要求. JR/T 0205-2020.

[7] The State of Software-Defined Storage, Hyperconverged and Cloud Storage.白皮书. DataCore.

[8] Replication.官方手册. MySQL.

[9] Semisynchronous Replication.官方手册. MySQL.

[10] Group Replication. 官方手册. MySQL.

[11] Windows Server 故障转移群集与 SQL Server.官方文档. SQL Server.

[12] 什么是 Always On 可用性组？.官方文档. SQL Server.

[13] Shao J, Yin B, Chen B, et al. Read Consistency in Distributed Database Based on DMVCC[C] IEEE, International Conference on High PERFORMANCE Computing. IEEE, 2017:142-151.

中国信息通信研究院 云计算与大数据研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：13691032906

传真：010-62304980

网址：www.caict.ac.cn