2024百易存储研究院 · 少数派报告（2）

“两地三中心”建设需求：分布式存储更有优势

百易存储研究院

DOIT传媒

2024年5月

前言

“两地三中心”方案最早出现在金融行业，源自国民经济重点行业对计算机系统 RTO（(Recovery Time Objective, 复原时间目标，简单说就是故障恢复需要花费的时间) 的极为苛刻的需求，哪怕 RTO 为 1 秒都是不可容忍的，都会带来巨大的影响和的损失。

为了避免灾难而导致业务中断，金融行业提出了“两地三中心”建设的需求，作为保障业务连续性的双保险。“两地三中心”中的两地是指同城、异地；三中心是指生产数据中心、同城容灾数据中心、异地容灾数据中心。在早期，“两地三中心”建设几乎成为了重点行业的标配，是刚需，也是行业监管部门的硬性要求。

当时光来到互联网时代，分布式技术异军突出，x86 通用计算、分布式存储开始扮演重要的角色，开始蚕食、替代小型机和专属存储设备的市场，在互联网领域取得成功。以互联网企业为样板，金融等重点行业用户也开始尝试使用分布式技术，但是他们也经常性的提出一个问题：分布式存储是不是也能够支持“两地三中心”建设的需求?

目录

CONTENTS

分布式存储“如鱼得水”

很多时候，我们可以把分布式存储作为一个完整产品形式来看待，无缝衔接、替换原有集中式存储设备，从这个意义上来说，分布式存储可以支持“两地三中心”。

但现实的情况是， 分布式存储支持“两地三中心”方式有所不同，也可以说更加具有效率。

分布式存储采用的不是“1+1+0.5”的方式，而是可以用 1.5 来表达的方式。原因也很简单，“两地三中心”本质是不是就是“分布式”？当“分布式”需求遇上“分布式（的）存储”，岂不如鱼得水？

所谓惊喜也由此得来！

分布式存储“两地三中心”方案设计

（这部分内容比较专业，涉及很多技术细节，用户可以根据需要，如果不关注技术，可以直接跳过本章节。）

为了更好展示分布式存储“两地三中心”方案的设计与技术特点，我们以一个典型的银行数据存储系统容灾需求为例，展开介绍。

场景介绍：

银行 A 是一家中大型金融机构，其业务包括个人银行业务、企业银行业务、投资银行业务和在线交易服务。银行的核心系统存储着关键的财务数据、客户信息、交易历史和合规记录。银行希望通过容灾方案保护关键数据资产，RPO 和 RTO 都要满足金融监管要求。

设计原则

（1）满足业务需求：整体设计以满足业务需求为第一目标，仔细识别各种业务对存储系统的需求，将指标量化，根据量化指标选择合适的方案。

（2）可靠性：可靠性的设计重点关注两个方面，一个方面是各组件本身的可靠性，一个是各组件整合在一起时，架构本身的可靠性。

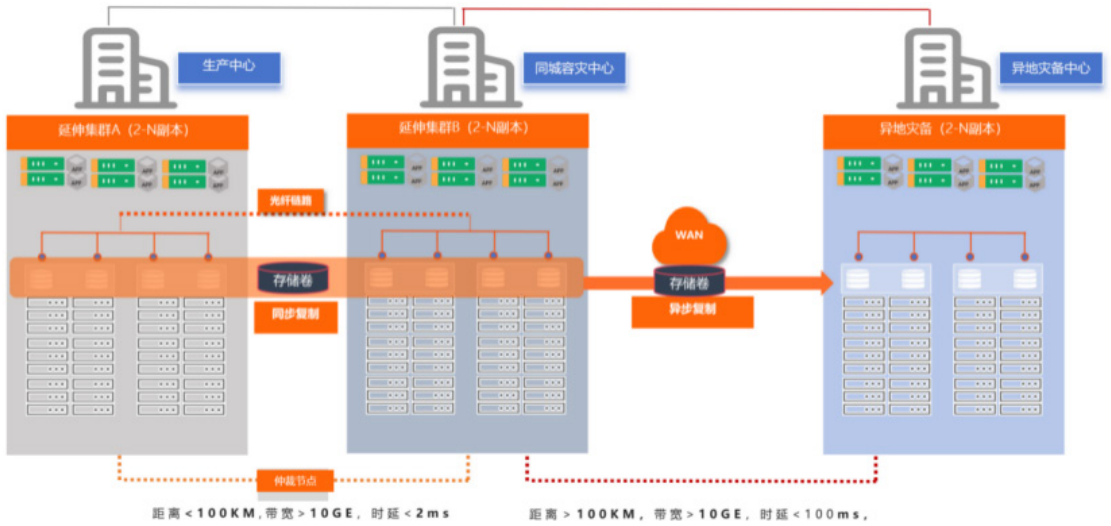
（3）性能：由于不同应用对存储资源的需求不同，在设计时将性能作为一个重要因素进行考虑，指导对于不同的业务系统进行存储资源的分配，以充分利用分布式存储的性能，满足业务对性能的要求。

（4）扩展性：设计时不仅要满足现状，还要考虑后续的扩展性，在扩展时整体框架保持稳定，降低架构调整对应用系统的影响。

（5）易维护性：产品后期维护也是需要考虑的一个关键方面。因为采取不同的设计可能造成以后维护的频

度和工作量不同。设计上尽量采用便于后期维护，维护成本低的方案。

2. 分布式存储两地三中心方案架构



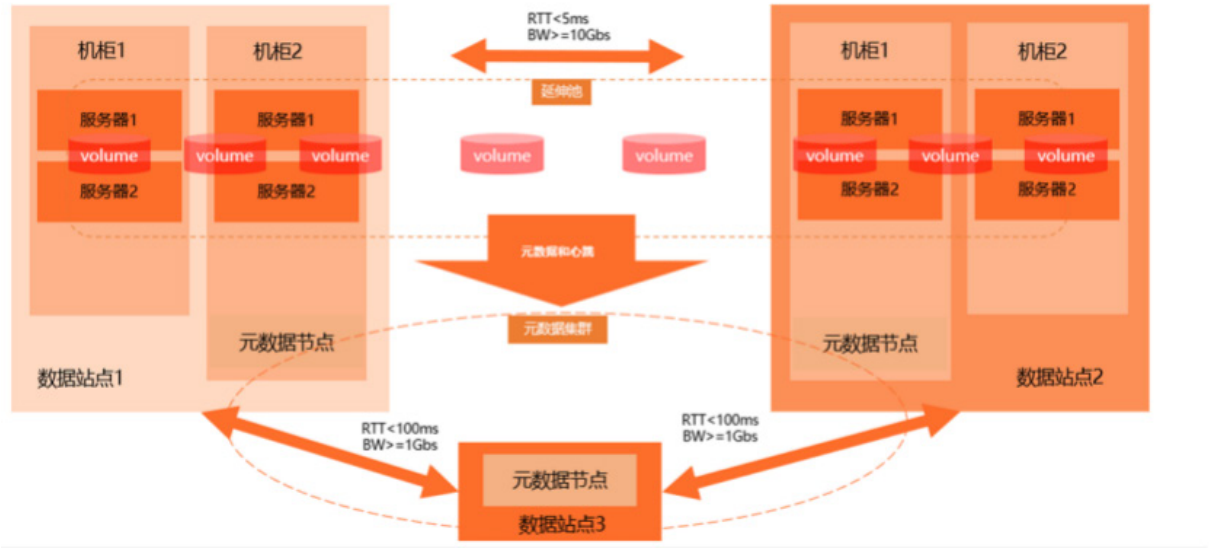
- 保护策略：数据以延伸集群跨副本方式把多副本数据存放在同城的两个数据中心，实时同步的方式实现同城容灾双活的需求；采用 Snapshot 方式本地站点数据与跨区域站点（任意距离）数据进行快照备份，实现异地灾备。
- 副本策略：延伸集群每个数据中心推荐 2 副本策略，以保护两个数据中心数据安全与业务连续性。
- 数据保护粒度：集群内配置多池策略，对不同应用创建不同存储池，存储池在不同数据中心设置不同主备策略，实现双中心业务就近读写；异地灾备采用 ROW 无损快照，通过分钟级、小时级等不同时间间隔快照策略，满足不同业务的数据保护需求。
- 容灾能力：同城容灾双活，实现 RPO=0，RTO ≈ 0（依赖业务切换时间），结合业务层可实现自动切换；异地灾备实现 RPO= 分钟级，RTO 依赖业务切换时间，存储实现分钟级挂载。

核心技术

（1）存储层同步远程复制

存储层的同步复制技术，其核心是利用存储系统自身提供的基于数据块的复制功能，将本地站点的数据实时的复制到远端站点。通常，同步远程复制要求两个站点的存储系统是同构的。在传统存储中大多基于卷级别的复制功能实现，而在分布式存储中，绝大多数都采用延伸集群的方式实现。

延伸池本质上是一个存储池在地域上的延伸，普通的存储池只能部署在同一个数据中心，而延伸池可以扩展到两个数据中心，通过副本分配算法，让存储池中数据副本跨站点分布，从而实现更高级别的可用性。同一个延伸集群中，可以创建多个延伸池，各延伸池的副本数和主副本位置都可以灵活设置。



- 延伸集群所需的物理资源条件

两个数据站点，每个数据站点至少需要 2 台服务器，以便构建 4 副本延伸池。

第三站点，该站点中部署一台元数据节点，元数据节点支持物理机或虚拟机。

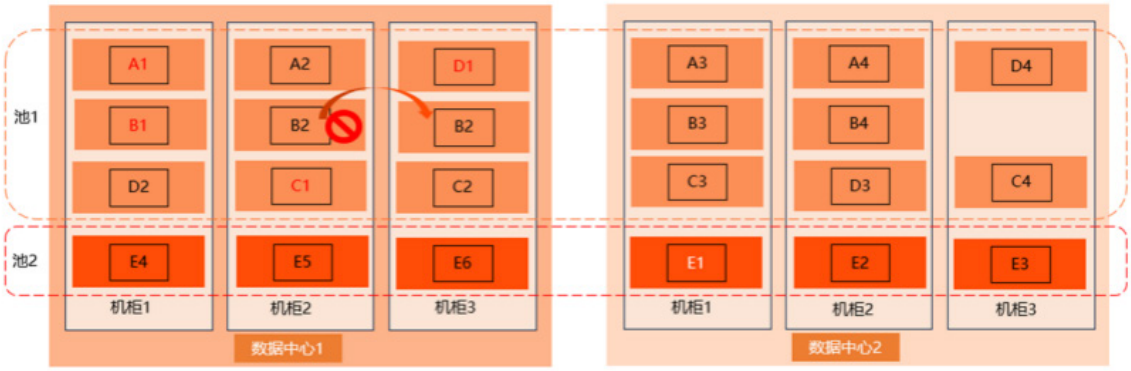
- 延伸集群所需的逻辑资源条件

延伸池：基于两个数据站点上的存储节点搭建延伸池，延伸池承载客户的 IO 业务。

- 延伸池的副本策略

支持 4，5，6 副本

支持设置优先读本站点的读策略



根据业务需要，可以创建不同冗余策略的延伸池，池 1 为 4 副本，池 2 为 6 副本

根据业务需要，主副本可以设置在不同的数据中心，保证就近写，池 1 主副本在数据中心 1，池 2 主副本在数据中心 2

- 延伸池的数据复制协议和故障处理

采用强一致性复制协议进行数据复制，并且支持从非主副本读取数据，从而实现就近读取

某一数据中心内硬盘或服务器故障时，数据重构只在当前数据中心内，减少跨数据中心的数据流动，节省网络带宽

数据中心整体故障之后，心跳机制可以快速检测到故障，从而通过广播通知客户端和另一数据中心硬盘管理进程进行视图切换，最终完成主备站点间切换

元数据集群：由两个数据站点和第三方站点上的至少 3 个节点，组成一个元数据集群。在元数据集群中保存了整个存储池的关键信息，例如存储节点的拓扑，副本复制关系（也称视图）等。同时，也承担了各个硬盘管理进程的心跳功能和主副本选举功能，以便在服务器或硬盘故障时，及时更新视图，保证业务连续性。元数据集群可以采用成熟的选举算法，例如 Bully 算法，Raft 算法，ZAB 算法。

- 延伸集群的性能

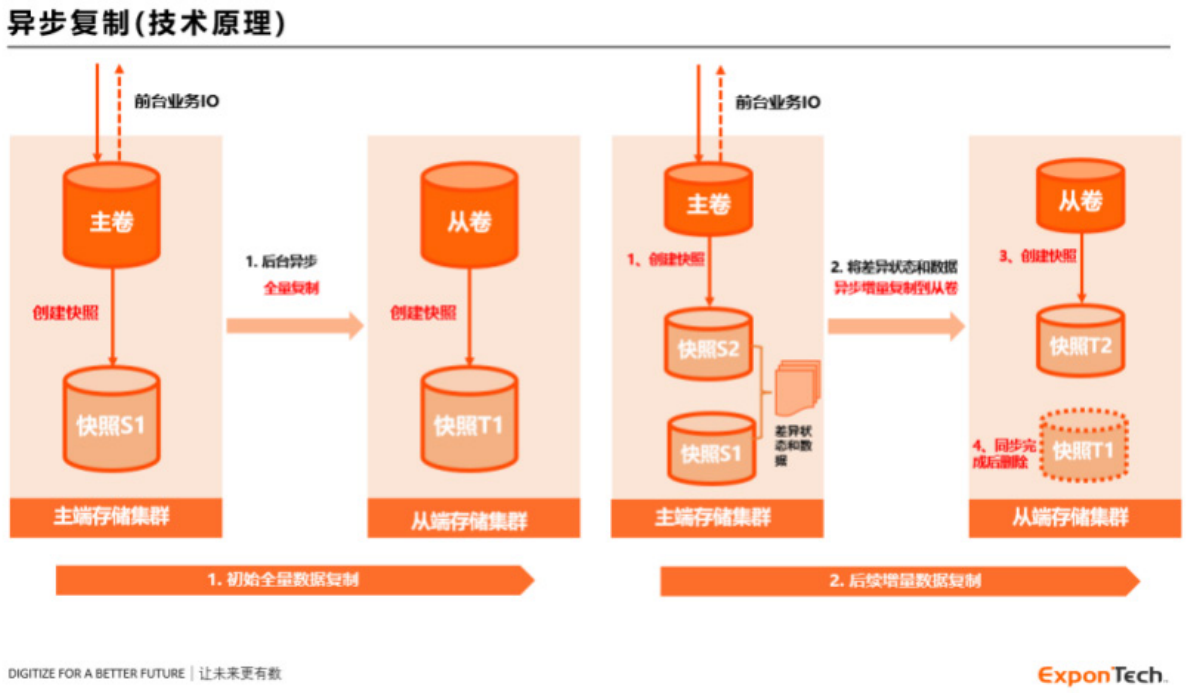
传统的分布式存储基于开源软件和 TCP/IP 网络构建，采用传统的系统软件模型处理硬盘和网络 IO, 总体上其 IO 栈很长，导致 IOPS 比较低，时延很大，这还只是在单站点内部署一套集群的情况。如果是跨越两个站点

部署延伸集群，则其时延就会更加恶化，很难满足关系型数据库等企业核心应用的要求，也很难在性能和端到端时延上与基于两套传统存储阵列构建的双活模式匹敌。

可喜的是，目前市场上已经出现了采用 SDS 2.0 架构的新一代分布式存储产品，能够在集群峰值压力下提供端到端 100 微秒以内的极低稳定时延，有了这个基础，即使构建一套跨数据中心的同步复制灾备集群，端到端的的时延也可以控制在合理范围内。因此，新一代 SDS2.0 分布式存储系统本身的极高 IO 性能和极低的稳定时延，在很大程度上弥补了两个数据中心之间的网络时延带来的性能损耗。

（2）存储层异步复制技术

异步远程复制是指将本地站点的数据周期性地同步到远端站点上。目前，业界实现异步复制技术主要分为基于日志机制和基于快照机制两类。其中，基于快照的技术在传统存储和分布式存储中都被广泛使用，其技术特点为:



- 按照卷的粒度对整个复制过程进行管理，通常需要将本地站点的卷和远端站点的卷配置 Pair 关系，本地站点的卷称为主卷，远端站点的卷称为从卷。

- 复制过程强依赖于快照技术。快照的性能很大程度上影响着异步复制的 RPO。
- 客户端对主卷进行写操作时，只要主卷返回写请求成功，即向客户端返回写请求成功，保证客户端的写入性能不受影响。
- 主从站点间配置 Pair 关系的卷，通过全量同步加增量同步的方式，保证主从站点两个卷之间的数据达到一致。

根据以上的技术特点，可知，异步远程复制技术中，最为关键的点包括：

- 快照本身的性能：在数据周期性同步过程中，存储系统要不断的打快照和删快照，两次快照之间的最小间隔，决定了差量数据的多少，最终决定了整个异步复制方案的 RPO。因此，分布式存储中，为了实现高性能快照，都会采用 ROW 快照。

根据 ROW 快照的实现原理，随着快照链的长度增加，读数据时存在一些元数据检索上的开销，但是，优秀的分布式存储产品会设计专门的机制来优化元数据的管理，从而实现极小的读性能损失。例如，根据业内一家做新一代软件定义存储 SDS2.0 厂商的性能数据，快照链长度即使为 128，读性能几乎没有下降；而快照的写性能，与快照链的长度无关，几乎没有下降。通过无损 ROW 快照技术可以实现最小化 RPO、维持系统高性能、更利于数据密集型应用等核心价值。

- 差量数据的计算：两次快照之间差异数据的计算效率，很大程度上也决定了整个复制机制的效率。为了快速计算差量数据，通常会采用 bitmap 来保存卷上各个数据块的变化状态，当数据有更新时先更新 bitmap，再写入数据。为了进一步提升效率，多级 bitmap 或更细粒度的 bitmap 等技术。
- 存储本身的性能：在数据同步过程中，会额外产生本地站点的读业务（读取差量数据），本地站点的写业务（删除快照产生数据 merge），远 44 端站点的写业务（写入差量数据），以及大量的元数据操作（频繁创删快照）。因此，分布式存储本身的 IO 读写性能越高，越容易缓解这些后台业务对前台业务的影响。
- 前后台业务流控：在数据同步过程中，需要一个很好的流控机制，才能保证数据同步过程中，前台业务不受影响，或将影响降到最低。

（3）存储高性能

两地三中心方案一般用于核心业务，追求数据的高可用和业务的极致稳定。这种部署模式通常用于关键业务的灾难恢复和持续运行，但为了提高数据的同步效率、减少 RPO 和 RTO、支持高并发访问等业务需求，存储的高性能成为必要的基础。从分布式存储技术的发展看，以下核心技术是不可或缺的。

- 高性能 IO 软件栈：采用 RTC（Run to completion）和全栈无锁化架构，IO 会按照 ID Sharding 到不同的通道并端到端下盘，端到端实现资源 Sharding 分割，实现全无锁。高性能 IO 软件栈在简化设计的同时可以有效提高性能、减少阻塞，极大提升数据处理的效率。
- 端到端零拷贝：它的目的是最小化 CPU 的数据处理负担，减少数据在系统中的拷贝次数，从而提高数据传输效率和整体系统性能。计算侧零拷贝，从协议侧接到 IO 申请内存到网络发送，全栈实现零拷贝。存储侧零拷贝，从网络接受到 IO 到下盘全栈实现零拷贝。
- 大页内存管理：在大页内存管理中，操作系统将多个连续的物理内存块预留给单个大页，而不是将这些块分散到多个标准页中。通过减少 TLB(Translation-Lookaside Buffer) 缺失 (TLB 发生缺失可以分为两种情况：如果该页在主存中，那么 TLB 缺失只是一次转换缺失；如果该页不在主存中，此时 CPU 就需要调用操作系统的异常处理。)
- RDMA 网络：基于 RDMA 网络，采用新一代网络软件组件以及新一代 IO 栈的分布式存储能够把单路访问时延降到 100us 以内的水平。目前市场上常见的 25G，100G 以太网交换机和网卡都已经支持 RoCE v2 协议，RDMA 网络已经具备在企业数据中心大规模使用的条件。

解读和结论

分布式存储与集中式存储实现“两地三中心”的方式有所不同，其中，集中式存储重点强调硬件的可靠连接，采取冗余设计的方案，因此可靠性、稳定性和可用性是非常高的，不存在单一故障点。对于“两地三中心”，集中式存储依靠数据复制，借助同城强大光纤连接通信能力，对数据进行保护，对系统进行容灾，很多时候，要求采用相同型号和配置的设备，并采用“双活”模式替代“主备模式”来提升效率。异步复制模式是两地不得已地选择，对设备型号和配置要求，也会有所降低。

对于分布式存储而言，满足“两地三中心”的需求，实质上就是副本数据跨数据中心的分散部署（延展集群方式），主要依靠快照同步 / 异步设计的思路。如以上提到的 ROW 无损快照、RTC、大页内存管理、RDMA 网络等，影响的是性能或者说效率。

在此，我们不难得出这样的结论，同样的“两地三中心”需求，分布式存储在实现技术上是有差别的？

分布式存储是不是更有效率？这就需要用户来仔细的品味了，总而言之，同样的“两地三中心”，实现的技术方法是不同的，只要注意到这一点，已经足够了！

百易存储研究院出品

2024 年 5 月

添加企微咨询

微信公众号