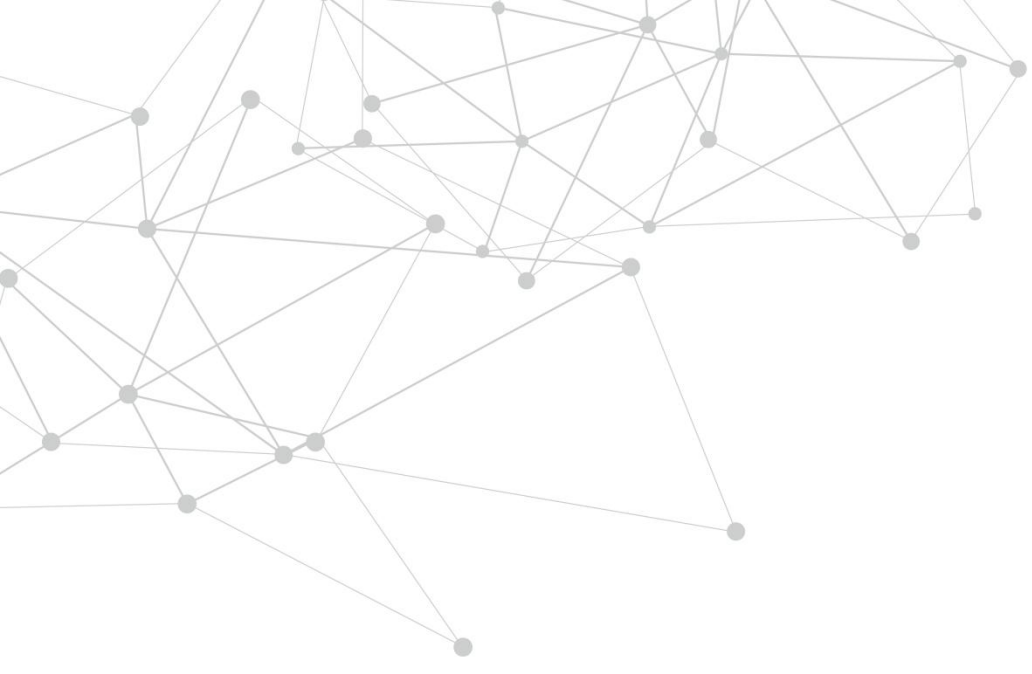




金融数据库存算分离架构 选型白皮书

北京金融信息化研究所（FITI）

2024 年 10 月



版权声明

本白皮书版权属于北京金融信息化研究所有限责任公司，并受法律保护。转载、编摘或利用其他方式使用本白皮书文字或观点的，应注明来源。违反上述声明者，将被追究相关法律责任。

编制委员会

主 任:

潘润红

副主任:

黄程林、庄文君

编委会成员（排名不分先后，按姓氏拼音排序）：

董晓杰、何东川、何佳佳、胡盼盼、黄 涛、简丽荣、王鹏冲、
杨传辉、张子鉴、朱 鹏

编写组成员（排名不分先后，按姓氏拼音排序）：

鲍思佳、陈灿荣、陈 辉、程 静、从平平、崔志伟、董 里、
金振山、李会亮、李云龙、李中原、梁 燕、廖美东、刘 常、
刘方彪、刘 昭、罗兰瑞婧、马 攀、马 涛、明玉琢、邵志杰、
盛 玉、万 里、王爱国、王 成、王帅强、王耀强、徐 赞、
杨 尚、张 刚、张 羿、赵 蒙

主编单位:

北京金融信息化研究所
中国农业银行股份有限公司
中国邮政储蓄银行股份有限公司
平安银行股份有限公司
上海银行股份有限公司
深圳前海微众银行股份有限公司
中国人寿财产保险股份有限公司
中国人民人寿保险股份有限公司
华为技术有限公司
北京奥星贝斯科技有限公司
武汉达梦数据库股份有限公司

参编单位:

北京酷克数据科技有限公司
阿里云（北京）科技有限公司
平凯星辰（北京）科技有限公司
中电科金仓（北京）科技股份有限公司
北京海量数据技术股份有限公司
天津南大通用数据技术股份有限公司
成都虚谷伟业科技有限公司
天津神舟通用数据技术有限公司

摘 要

金融业作为信息技术应用的前沿阵地，展现出对新技术的较高敏锐度和快速响应能力。中国人民银行印发的《金融科技发展规划（2022-2025 年）》中，强调打造数字绿色服务体系，建设绿色高可用数据中心，架设安全泛在的金融网络，布局先进高效的算力体系，夯实数字基础底座，助力数字金融高质量发展。

在创新发展和安全可控的双重驱动下，我国数据库作为金融信息系统的重要组成部分，在金融机构中的持续创新应用有效推动数字金融安全生态的构建。随着金融业数字化转型的深入，金融数据库架构正逐步从传统封闭走向创新开放。在此过程中，数据库存算一体架构可以有效满足金融新兴业务初期对快速和敏捷交付的需求，但随着金融业务的增长和基础设施规模的扩大，存算一体架构在可靠性和扩展性方面的局限性逐渐显现。因此，开放解耦的存算分离架构，不仅能够根据业务需求实现计算和存储资源的独立扩展，提供灵活的弹性扩缩容计算能力，还能利用存储的高可靠性优势，进一步提升整个集群处理海量数据的可靠性，重新成为数据库场景的架构选择。

本报告针对金融业务场景的特点和需求，深入剖析当前主流数据库存算分离技术架构方案，总结数据库存算分离架构选型评估指标体系，构建数据库存算分离架构实施路线，并总结金融机构的探索实践经验，为推动金融业数据库架构创新发展贡献力量。

目 录

1. 金融业数据库技术架构演变历程	1
2. 金融业应用对数据库的需求现状	3
2.1 金融场景业务特性	3
2.2 金融业数据库应用的基础架构问题和需求	5
3. 主流数据库存算分离架构技术分析	13
3.1 集中式数据库存算分离技术	14
3.2 分布式数据库存算分离技术	16
3.3 云原生数据库存算分离技术	22
4. 数据库存算分离架构选型评估指标与建议	28
4.1 性能指标	28
4.2 可靠性指标	29
4.3 可用性指标	31
4.4 成本效益分析	34
4.5 能耗指标	35
4.6 扩展性指标	37
4.7 安全性要求	38
4.8 运维指标	39
4.9 演进能力	41
4.10 总结	41
5. 数据库存算分离架构演进方案	42
5.1 外置存储的存算分离架构	43

5.2 存算协同的存算分离架构	44
6. 金融业数据库存算分离架构应用实践	46
6.1 工商银行数据库存算分离架构实践	46
6.2 农业银行新一代存算分离数据库实践	54
6.3 平安银行 MySQL 数据库存算分离架构探索	57
6.4 上海银行数据库存算分离架构的应用试点实践	61
6.5 微众银行 TDSQL + OceanDisk 存算分离架构	64
6.6 人保寿险数据库存算分离架构转型升级	71
7. 总结	76

1. 金融业数据库技术架构演变历程

金融业的数据库技术架构演变紧跟金融业务的发展步伐，同时与数据库和存储等技术的进步相互促进，共同推动金融业创新发展。在过去的半个世纪里，数据库技术架构的演变主要经历了以下三个阶段：

第一阶段（1960 到 1970 年代）：采用存算分离架构的文件系统数据库，支持单体应用。

早期的数据管理系统主要以文件系统为基础，先后发展出网状数据库（IDS）和层次数据库（IMS）等形态。IBM 于 1966 年推出的层次型数据库 IMS，与该公司的服务器、存储和中间件等产品紧密结合，提供了一站式解决方案，有效满足核心业务交易和数据处理对高可用性的需求，被广泛应用于金融、航空和制造等领域。即便在互联网飞速发展的时代，这些行业依然在使用 IBM 的 IMS 数据库。然而，由于单一厂商的封闭系统带来昂贵的服务费用和有限扩展能力，这种解决方案成为仅在金融等少数行业应用。

第二阶段（1970 年到 2000 年左右）：关系型数据库蓬勃发展，存算分离成为基础架构。

1970 年代，IBM 率先提出关系数据库模型并发布了 SQL 标准。1979 年，Oracle 公司推出了首个商用关系型数据库，而 IBM 直到 1983 年才发布其关系型数据库 DB2。此后，Sybase、SQL Server 等其他商用关系型数据库也陆续进入市场，开启关系型数据库的繁荣时代。尽管如此，Oracle 和 DB2 依然是市场最主流的商用

数据库产品。

与此同时，RAID 磁盘专利和 SCSI 接口标准的发布，以及 EMC 推出的高端存储产品，都在吞吐性能、系统可靠性和容量扩展性等方面实现显著提升。由此，IT 产业逐步形成以 IBM 小型机、Oracle 数据库和 EMC 存储为核心的 IOE 架构。这一架构成为“开放”和“解耦”的存算分离架构的基石，推动金融业信息化应用的发展，并促进金融业集中业务处理的区域性数据中心建设。

第三阶段（2000 年至今）：数据库技术呈现多元化发展趋势，存算分离架构依然是重要的架构方向。

2000 年前后，随着互联网技术的广泛应用，数据库技术进入多元化混合发展的时代，数据库应用更加多样化。在此期间，X86 服务器凭借其价格优势和通用性，一定程度上扩展了传统 IOE 架构的“开放性”，但由于 x86 服务器在吞吐性能上的局限性以及可靠性方面的不足，相较于传统 IOE 架构的服务器，新兴数据库应用被迫采用一主多副本、分库分片等存算一体的架构形态，以提升吞吐性能和可靠性。在金融互联网应用的初期，存算一体架构促进了互联网技术在金融业务场景中的迅速推广，成为金融业技术架构的重要补充。

但是，在存算一体架构下，上层应用对数据库不再是单一的数据访问，还需要考虑数据分布和可靠性等需求。这导致数据库架构设计对应用的侵入性增强，应用与数据库之间深度耦合，进而加剧了架构的“封闭性”。同时，随着金融新兴业务的扩张和基础设施规模的增大，存算一体架构在可靠性和扩展性方面的局限性逐渐显现，比如：存算一体架构中任意单节点的计算或存储

故障都可能引发数据丢失和业务中断，带来业务连续性风险；存算一体架构无法实现存储和计算资源的独立扩展，两者必须同步扩展，造成一定的资源浪费。

另外，随着 NVMe Over Fabric、RDMA 和存储介质等技术的成熟应用，计算与存储之间的网络带宽得到显著提升，有效缓解了存算分离架构在吞吐性能上的瓶颈。通过利用存储的 RAID 和快照等高可靠特性，可以设计出多层次的容灾和备份方案，既简化了系统架构和运维，又提高了整体业务系统的可靠性。

2. 金融业应用对数据库的需求现状

2.1 金融场景业务特性

金融业的业务场景多种多样，根据面向客户和内部管理的不同需求，可将业务系统主要分为两大类：生产交易系统（TP）和数据智能分析系统（AP）。

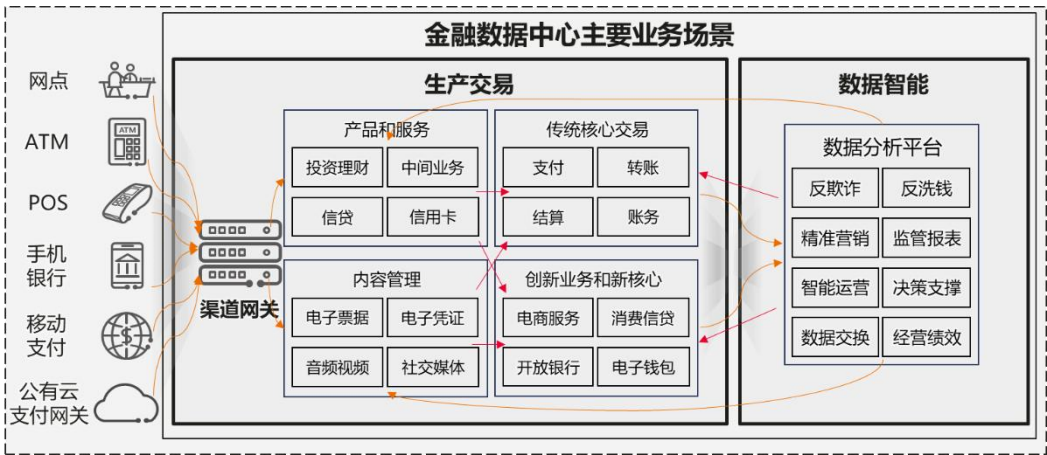


图 1 金融业主要业务场景

生产交易系统主要为客户提供一系列对外服务，覆盖支付与清算、核心交易、中间业务、银行卡、信贷管理、用户管理以及

渠道接入等关键领域。这类系统对实时数据处理、随机数据访问和业务连续性有着极高要求，因此通常采用本地双活加异地主备的容灾方案。

鉴于生产交易系统的联机事务处理（OLTP）特性，OLTP 数据库因其出色的实时响应能力、高并发处理能力、强数据一致性以及高可用性等优点而受到青睐。根据金融信息化研究所《金融业数据库创新发展报告（2024）》数据显示，在金融业的数据库应用领域中，OLTP 数据库实例数占比 65%，充分体现了其在维护金融服务稳定性方面的核心地位。

数据智能分析系统虽不直接面向客户，但在辅助银行实现高效运营和管理方面发挥着重要作用。该平台支持多种业务系统，如反欺诈和反洗钱等风险管理系统、银行产品绩效管理、用户画像和精准营销等。

数据智能分析系统正逐步演进为支持准实时处理的系统，为关键业务提供及时的决策支持。这类系统具备批量数据处理和顺序数据访问的特点，因此容灾方案主要采用同城或者异地的主备容灾方案，鲜有本地双活方案。

在处理联机分析处理（OLAP）业务的数据智能分析平台中，OLAP 数据库凭借其对复杂查询的高效处理、灵活的数据分析能力以及支持多维数据模型等优势，满足数据智能分析平台对数据处理和分析的需求。在金融业数据库应用中，OLAP 数据库实例数占比约为 21%。

近年来，随着大数据技术的迅猛发展，金融数据分析平台的数据湖和湖仓一体正逐步转向存算分离架构。然而，在生产交易场景中，数据库处理的业务不仅至关重要，而且结构极为复杂，使得从传统封闭架构向开放解耦架构转型的过程中面临诸多挑战，包括性能、可靠性、扩展性、成本和能耗等多个方面。因此，本文将着重分析生产交易场景下数据库处理的业务特性。

2.2 金融业数据库应用的基础架构问题和需求

随着金融数字化转型的深入，金融业务形态发生显著变化，特别是面向客户提供服务的网上银行、手机银行和支付网关等新兴技术的应用，使得金融业务模式从传统的线下网点服务扩展到线上服务，从原来的 5*8 小时服务延长到 7*24 小时不间断业务服务。这些业务模式的变化，极大地推动了移动金融服务的快速发展，也正在深刻影响并重塑着金融业数据库应用的基础架构。

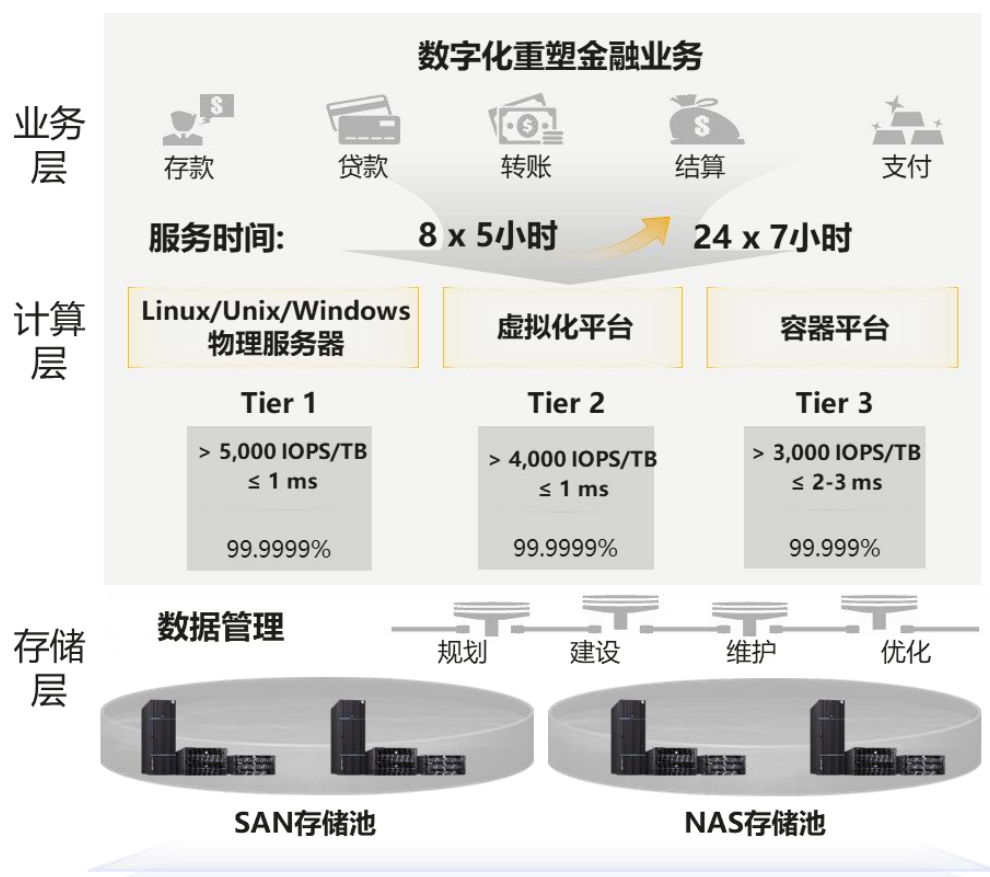


图 2 金融业生产交易场景的基础架构

从多数银行公布的年报来看，新兴金融服务交易量年均增长率超过 50%，数据量年增长率约为 30%。新兴业务模式拓宽金融服务的触达渠道，延长服务时间，同时也正在重塑金融业的数据中心基础架构。为适应未来业务发展，数据中心基础设施必须在性能、可靠性、扩展性、能耗和总成本之间寻求最优架构。具体表现如下：

2.2.1 吞吐性能

业务量的激增对数据基础设施的性能和吞吐量提出更高要求，例如数据库读写时延需小于 1ms。一般银行业务平均 SQL 请求数约为 30-50 次，每次 SQL 请求涉及约 10 次以上的存储 IO 读

写。以中等银行生产业务峰值交易量约 5000 笔/秒（即 5000TPS）为例，峰值业务处理所需的存储 IO 请求能力约为每秒 1.5–2.5 百万次读写（IOPS）。这种吞吐性能对传统存算一体架构的服务器构成巨大挑战，因为服务器 CPU 算力不仅要负责业务处理，还要完成数据可靠性、数据加密和数据压缩等复杂的存储处理任务，导致 CPU 资源紧张。然而，通过采用存算分离架构，所有数据存储的 IO 请求可以卸载到外置存储系统中处理，从而减轻服务器 CPU 的负载，使其无需处理存储 IO 请求，进而提升服务器 CPU 在业务处理方面的能力。

2.2.2 可靠性

服务时间的延长对整体架构的可用性提出更高要求，数据中心基础架构的可靠性需提升至 99.9999% 以上，以满足整体业务可用性 99.999% 的要求。

中等规模银行的核心系统交易量通常在 1000–3000 笔每秒，每秒的系统中断和不可用都会导致 1000 笔以上的交易损失和用户流失。存算一体架构的 x86 服务器在可靠性上仅能达到 99%。金融业统计显示，x86 服务器使用超过 5 年，故障率超过 0.5%。如果采用服务器本地磁盘作为数据库存储，服务器上的 RAID 卡或硬盘亚健康等故障都可能影响数据库的数据访问，给生产交易系统的业务连续性带来严峻挑战。而在存算分离架构下，服务器仅负责数据逻辑处理，数据的持久化和可靠性等能力则由存储设

备统一承担，且存储设备的可靠性远高于服务器，通过存储阵列 RAID 技术和快照技术等特性，可以极大提升整体业务的可靠性。

服务器本地盘通过 RAID 卡串联多块硬盘组成 RAID 组，主要用于在数据故障后提供冗余保护。然而在可靠性方面，RAID 卡本身存在单点故障的风险。

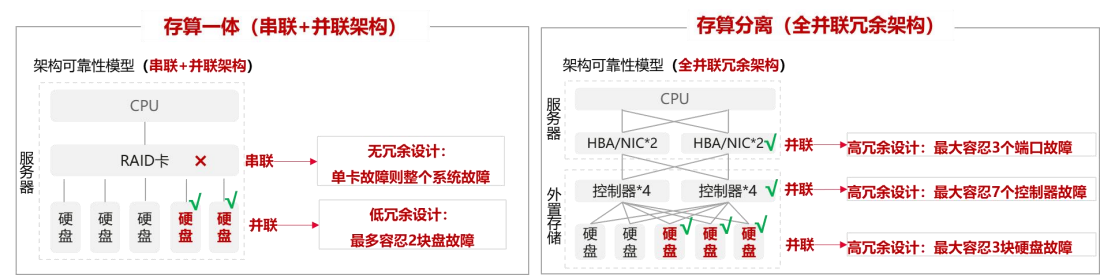


图 3 存算一体和存算分离架构的存储冗余设计对比图

通过对两种系统形态的详细测算，可以看出本地盘数据可用性仅为 3 个 9，而存算分离形态采用了全并行冗余架构，数据可用性可达到 5 个 9，实现 2 个数量级的提升。

	系统可用性 = RAID卡可用性 × 硬盘可用性 = 99.9% （年平均中断时长520分钟）	系统可用性 = HBA/NIC可用性 × 外置存储可用性 = 99.999% （年平均中断时长5分钟）
MTBF	单盘年故障率 = 0.4%，双盘同时故障率 ≈ 0.0016%，RAID卡故障率 = 0.001% $MTBF(RAID卡) = \frac{1}{RAID卡故障率} = \frac{1}{0.00001} = 100,000 \text{小时}$ $MTBF(双盘) = \frac{1}{双盘故障率} = \frac{1}{0.000016} = 62,500 \text{小时}$	HBA/NIC卡年故障率 = 0.05%，HBA/NIC双卡同时故障率 = 0.000025% $MTBF(HBA/NIC双卡) = \frac{1}{HBA/NIC双卡同时故障率} = \frac{1}{0.00000025} = 4,000,000 \text{小时}$
MTTR	MTTR(硬盘) = 每块盘(4TB)重构时长 > 20小时 MTTR(RAID卡) = RAID卡维修时长 > 2小时	MTTR(HBA/NIC) = HBA/NIC卡维修时长 > 2小时
可用性	可用性(HBA卡/NIC) = $\frac{MTBF(RAID卡)}{MTBF(RAID卡) + MTTR(RAID卡)} = \frac{100000}{100000 + 2} = 99.998\%$ 可用性(硬盘) = $\frac{MTBF(双盘)}{MTBF(双盘) + 2 \times MTTR(硬盘)} = \frac{62500}{62500 + 40} = 99.93\%$ 系统可用性 = 可用性(硬盘) × 可用性(RAID卡) = 99.93% × 99.998% = 99.9%	可用性(HBA卡/NIC) = $\frac{MTBF(HBA/NIC双卡)}{MTBF(HBA/NIC双卡) + MTTR(HBA/NIC)} = \frac{4000000}{4000000 + 2} = 99.99995\%$ 可用性(外置存储) = 99.9999% (RAID2.0、全互联架构、8坏7、硬盘亚健康管理等) 系统可用性 = 可用性(HBA卡/NIC) × 可用性(外置存储) = 99.9999% × 99.9999% = 99.999%

图 4 存算一体和存算分离架构可靠性对比图

2.2.3 业务连续性

金融业对系统业务连续性有着严格要求，其中数据库集群尤为重要，用户通常按照灾备建设的两大标准和两大衡量指标来规划总体容灾架构。

中国标准 《信息系统灾难恢复规范》	国际标准 英国Share78	RPO	RTO	等保2.0 技术要求
Tier 6-0数据丢失, 自动拉起业务	Tier 7-0数据丢失, 自动拉起业务	0	< 15 min	四级 本地备份+异地备份+本地高可用+异地业务高可用
Tier 5-0数据丢失	Tier 6-0数据丢失	0~30min	< 2 Hr	
Tier 4-电子传输和完整设备支持	Tier 5-灾备中心有实时的状态更新	2~12Hr	< 24 Hr	三级 本地备份+异地实时备份+本地业务高可用
	Tier 4-活动状态的灾备中心	2~24Hr	< 24 Hr	
Tier 3-电子传输和部分设备支持	Tier 3-电子链接	12Hr~24Hr	24 Hr	二级 本地备份+异地定时备份
Tier 2-备用场地支持	Tier 2-PTAM卡车运送+热备份中心	24 Hr~数天	24 Hr~数天	
Tier 1-基本支持	Tier 1-PTAM卡车运送	数天	数天	一级 本地备份
	Tier 0-No off-site Data	数天~?	数天~?	

图 5 中国和国际的信息系统灾备恢复标准的对应关系图

例如，对于 5-6 级核心类业务，除了要求数据本身的高可用外，对业务连续性也有极高要求。业界普遍做法是采用存储本地双活或存储同步复制方案，结合数据库层的日志异步复制方案，如 Oracle RAC + Oracle ADG。该方案可以保障同城跨数据中心双集群实现 RPO=0，即数据无丢失，并通过叠加异地容灾技术，实现两地三中心分钟级 RTO 恢复时间的容灾能力。

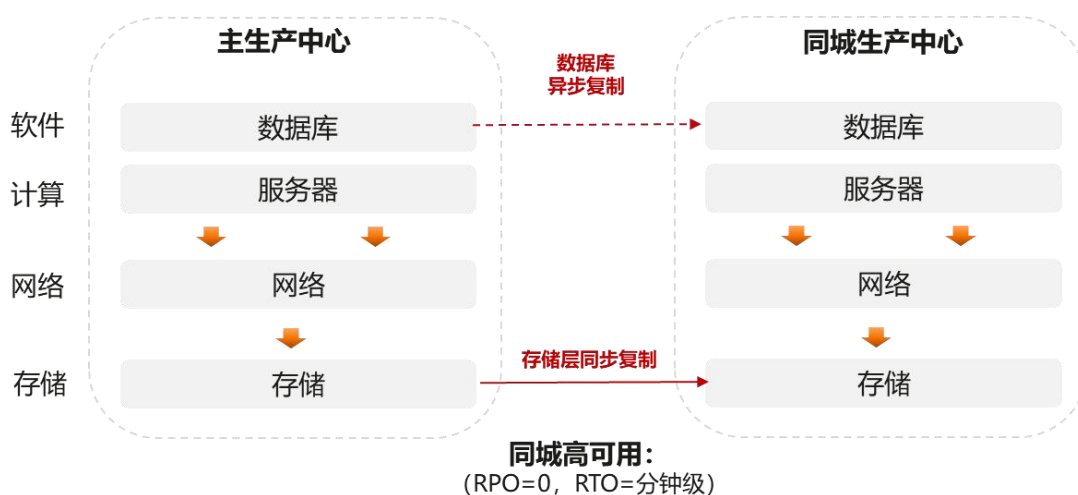


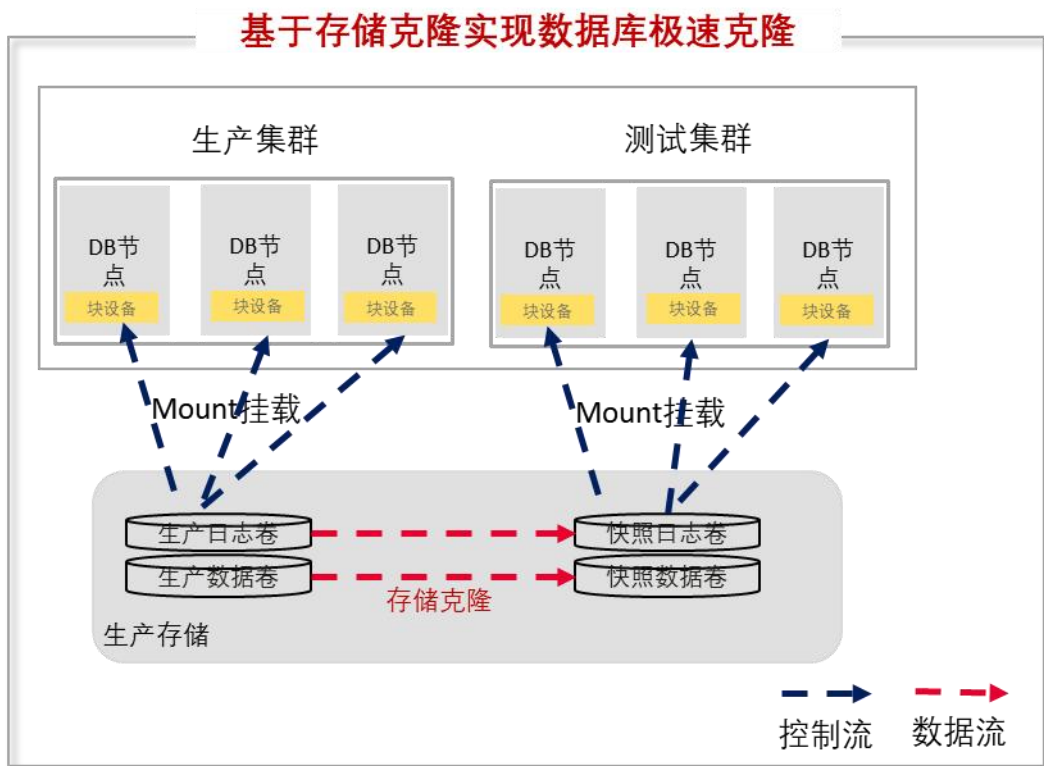
图 6 存储同步复制和数据库异步复制方案示意图

2.2.4 扩展性

随着业务种类的增加，金融机构对数据库的需求变得更为多

样，对各类运营数据和报表系统等从生产系统数据库中获取数据的实时性要求也随之提高。

银行生产交易系统作为各类系统运营数据和报表等系统的数据源，如何实时获取有效数据的同时最小化对生产系统业务的影响，成为各业务系统与生产交易系统之间的一大挑战。在存算一体架构下，通常采用专用 ETL 工具抓取数据，会对业务系统造成性能干扰，因此只能选择在夜间业务低谷时段抽取数据，这对后端系统分析数据的实时性造成影响。另外，每日批量作业需要在每日夜间定时从主库快速生成一个数据库集群，用于执行每日的读写操作。在存算分离架构下，可充分利用存储的一致性快照和克隆等能力，快速构建生产交易系统的数据库副本，满足各类后端系统对业务数据实时性需求，并且对生产端数据库的性能影响最小。



证券机构在验证系统变更后的业务连续性时，希望能在周末对主库进行快照，用于通关测试。测试完成后，通过整库闪回使数据库恢复到测试前的状态。

在存算分离架构下，可以充分利用存储快照和闪回特性，快速实现对源库数据和状态的保护及快速恢复，具有较高的闪回效率和最小的业务影响。

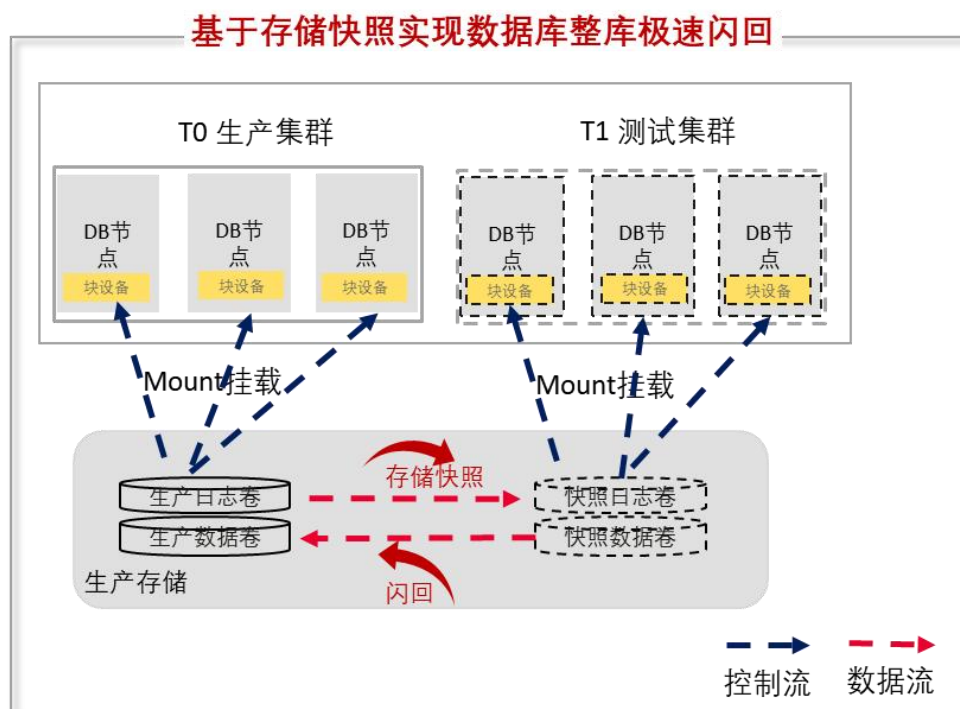


图 8 基于存储快照技术加速数据库闪回示意图

另一方面，越来越多的数据库集群在日常作业中涉及数据库备份、每日批量作业或周末通关测试。为解决大容量数据库场景下的快速备份问题，通过数据库和存储协同，实现 LAN Free 备份，提升备份速度，缩短备份窗口时间并减少对业务的影响。

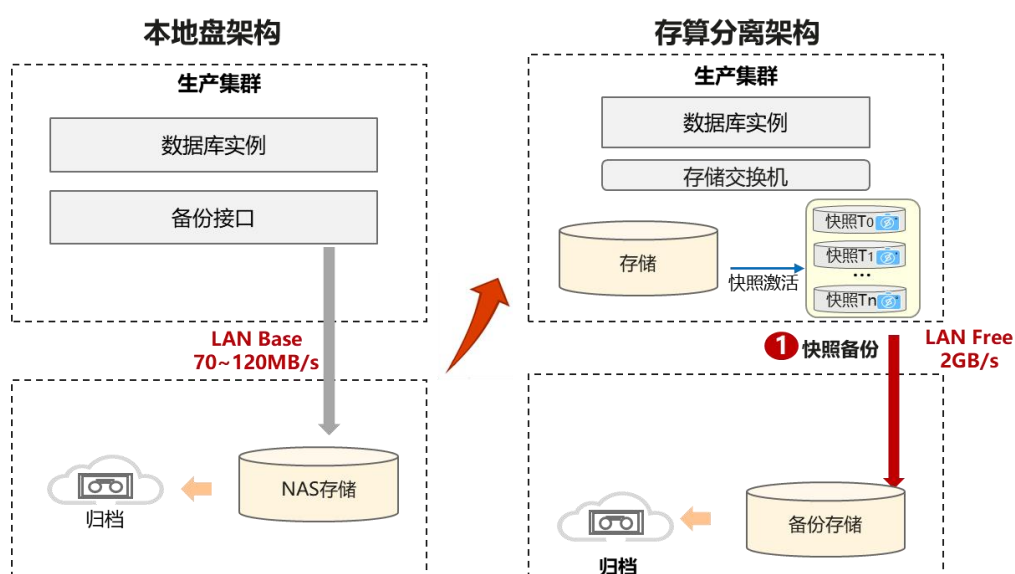


图 9 存算分离架构和存算一体架构下的备份方案对比图

2.2.5 系统运维

多样化的技术平台（虚拟化和容器平台）承载着多样化的业务，产生更加多样化的数据（包括结构化及非结构化数据），这使得各平台的运维复杂度不断增加。

银行业务种类繁多，各类应用平台混合存在，因此数据库种类多样，对存储的诉求也各不相同。在存算一体架构下，随着集群节点的增加，各类数据库的容灾、备份和日常运维等工作量呈指数级增长。同时，数据中心的空间占用和耗电量也急剧增加，导致大多数银行的数据中心面临新建或扩建的需求。

在存算分离架构下，采用分层解耦、按需分配的统一资源池方式，可以根据业务系统的等级分配不同的计算和存储资源，并通过 Quota 配额和 QoS 服务质量实现应用的隔离和资源保障。存算分离架构能够兼顾物理服务器计算资源、虚拟化平台和容器平台的多样性数据读取需求，提高整体资源利用率和运维效率。此外，随着全闪存储的普及应用，存算分离架构下的全闪存储在数据中心空间使用和能耗等方面进一步降低数据中心总体成本。

3. 主流数据库存算分离架构技术分析

随着数字化转型的加速推进，业务形态正经历深刻变革，传统应用与云及互联网业务场景日益融合，业务动态性显著增强。这不仅对基础设施资源的灵活性和利用率提出更高要求，还伴随着数据量的指数级增长和数据类型的多元化，给数据库带来实现

高性能和高可靠的挑战。与此同时，业务高峰的持续上升还要求数据库具备出色的弹性伸缩能力、持续服务能力以及合理的成本。

在多重因素的驱动下，无论是云原生、分布式还是传统集中式数据库市场，都在积极探索存算分离领域，数据库部署架构呈现从“存算一体”向“存算分离”的演进趋势。

3.1 集中式数据库存算分离技术

集中式数据库的核心特征在于数据的集中存储，这种架构为用户提供了与单机数据库相似的使用体验，数据在逻辑上可以统一访问，从而简化应用开发和运维工作，无需过多考虑数据分片、分布式事务等问题，更好地支持存储过程、多表关联和复杂查询。我国集中式数据库主流架构可分为主备多副本和存算分离。据金融信息化研究所《金融业数据库供应链安全发展报告》（2022）的数据显示，在金融业中集中式数据库占据主导地位，整体占比达到 89%。具体来看，银行业中的占比接近 79.62%，证券业和保险业的占比均超过 90%。详细情况如下图所示。

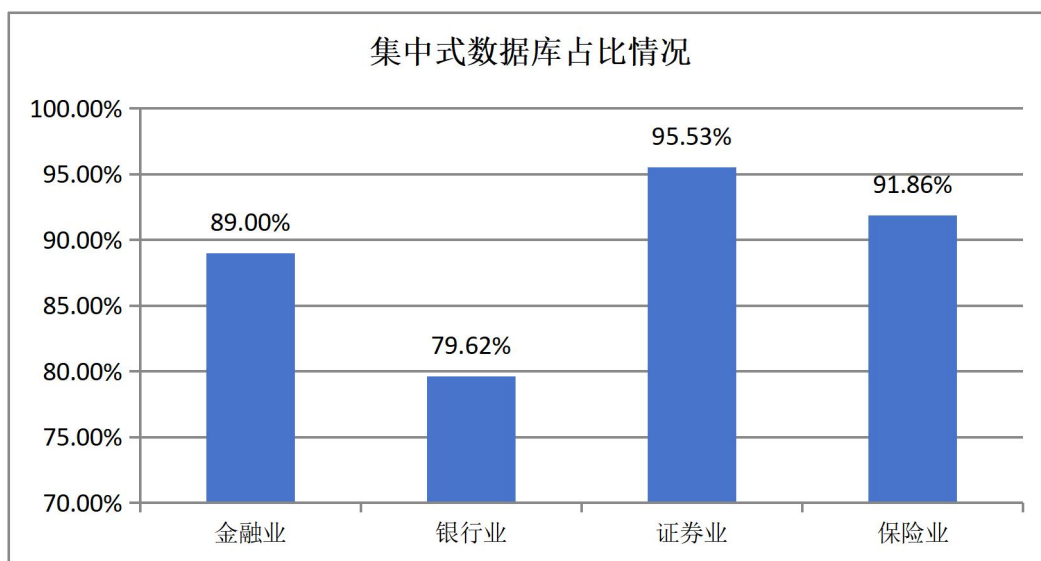


图 10 金融业集中式数据库占比情况示意图

集中式数据库凭借其较强的功能黏性、优秀的系统稳定性和良好的软硬适配能力，在金融业的应用仍占据绝大多数份额。随着金融业数字化转型的不断深入，集中式数据库需要更有效地支持海量、高并发、高吞吐量的新型金融业务应用系统。实现这一目标的关键在于提升专业共享存储能力，并结合高带宽、低延迟网络优化数据库架构。集中式数据库可以采用存算分离的部署方式，以适应不断变化的金融业务需求。

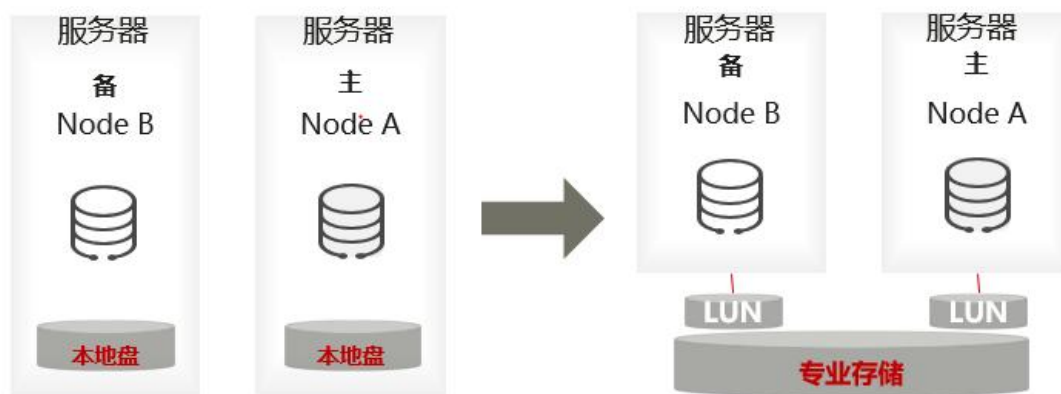


图 11 集中式数据库从本地盘部署到存算分离部署架构图

目前，国内基于主从多副本架构的集中式数据库由于实现简单、技术门槛低，应用较为广泛。然而，这种架构的数据库主要存在如下瓶颈：一是受制于服务器可靠性导致存算一体架构爆炸半径大、本地盘数据重构时间长；二是存算资源无法灵活扩展导致主从多副本架构需要较多硬件冗余，资源利用率较低，故障后数据库服务恢复速度较慢；三是由于日志同步过程中产生的性能损耗较大，特别是有大量变更或大事务时，采用同步复制会明显影响性能，使得一些主从多副本数据库产品在高可用设计中不得不在性能和数据一致性之间做出权衡。

采用存算分离架构的集中式数据库能够有效解决以下关键问题：一是使用可靠性更高的专用存储设备提升数据持久化可靠性，同时减少硬盘故障及故障恢复重构对数据库服务的影响；二是通过存储的高可用能力，替代传统的数据库日志复制方式进行持久化和容灾，减小对性能的影响。存算解耦使得计算和存储资源能够按需分配，提高资源利用率；三是随着存储能力的提升，存算分离架构允许数据库进一步实现副本缩减、容灾、备份能力下沉，以及 I/O 缩减和性能加速。通过采用控制器等针对数据库优化的专业存储设备，可以有效提升数据库性能，加速数据库密集型负载的处理。

3.2 分布式数据库存算分离技术

分布式数据库在金融业主要应用于敏态业务场景，其核心特

征是数据跨多个节点分散存储，并通过分布式事务实现并行处理，从而提升数据库的并发性能和容量。相较于集中式数据库，分布式数据库有效解决横向扩展的难题，但由于分片间的数据同步和汇总依赖于网络通信，尤其是在维护分布式事务的强一致性时，性能开销显著，因此更适合一些数据能够完美分片的业务场景。

当面临大量分布式事务或复杂的多表间关联查询时，分布式数据库的性能可能会大幅下降，进而影响其线性扩展。据《金融业数据库供应链安全发展报告》显示，我国分布式数据库在银行业使用比例仅为 17.51%，而证券和保险行业这一比例甚至不足 4%。究其原因我国数据库研发周期短且应用范围有限，其产品能力尚不完善。随着分布式数据库在核心业务不断深入应用，当前多数国产数据库采用存算一体、一写多读的架构，并通过单集群拉远的方式来实现容灾，导致在数据库升级过程中频繁遇到一系列问题。

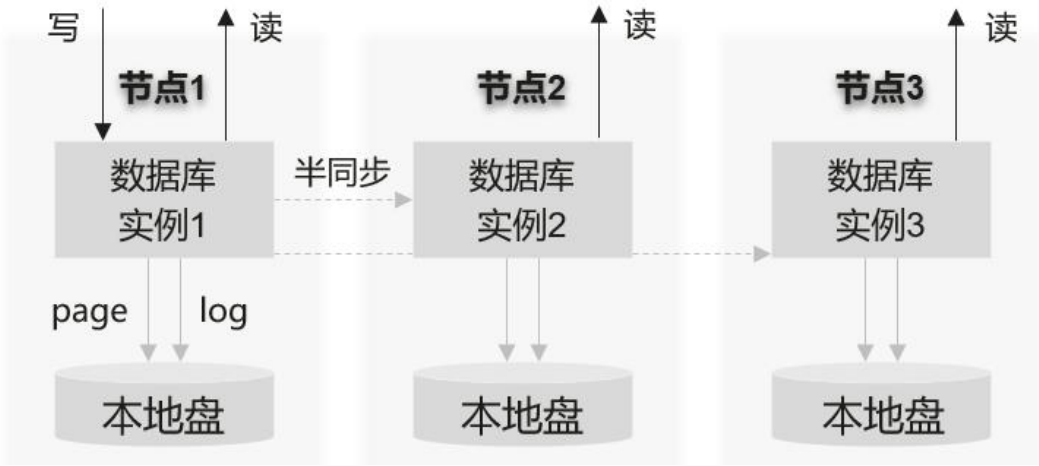


图 12 数据库存算一体部署架构图

3.2.1 业务改造难度大，风险高

在进行我国数据库改造时，需要通过分库分表的方式，将大型数据库拆分为多个小型数据库，以提高系统的并发访问能力和容量。然而，这种方法也带来了一系列挑战。

一是复杂性增加。业务系统改造涉及数据拆分、数据路由和数据同步等方面工作，这增加系统的复杂性和实施难度。例如，一家头部股份制银行在升级改造中，分库数量超过 15 个，分表数量超过 500 张，大幅增加了运维难度，甚至需要专门开发运维平台来应对系统复杂度的提升。

二是分布式事务影响性能。我国数据库在处理跨节点分布式事务时，可能导致数据库性能下降。例如，北京某银行在未进行业务改造前测试一款分布式数据库时，发现因分布式事务导致性能下降了 50%。

三是管理和维护困难。改造后的系统需要更多服务器节点来支持运行，增加故障定位定界的复杂度，导致运维投入显著增加。以某全国性银行的信用卡数据库系统为例，原系统仅需 4 台小型机，而分布式改造后采用服务器本地盘方式部署则需要 120 台数据库服务器，运维人员增加了 66%，开发人员增加了 5 倍，5 年整体拥有成本比原来提高了 60%以上。

综上所述，当前数据库转型升级面临诸多挑战，包括改造难度大、涉及面广以及管理和维护难度的增加，都可能带来重大的业务风险。

3.2.2 双碳集约目标面临挑战

数据库分库分表及或多副本冗余机制会导致集群规模的扩大和服务器数量的激增，同时服务器资源利用率却极低，进而造成投资浪费、能耗增加、运维困难以及机房空间不足等问题。以某省电信为例，数据库改造后 CPU 利用率仅有 5%，数据库服务器数量增加逾 10 倍，新增年功耗 22 万千瓦时，增幅高达 65%，显然无法满足企业降本增效的需求。某国有大行数据库改造后，存储资源利用率仅 3%，CPU 利用率仅不到 10%，数据库服务器数量更是达到了 20000 多台，迫使该行因机房空间不足而需扩建数据中心。这种大规模的数据库服务器部署和极低的资源利用率，对金融业践行双碳集约目标、实现成本精细化控制以及推进绿色节能的数据中心建设带来较大挑战。

3.2.3 系统可靠性下降

服务器本地盘的健康状态关乎数据库的稳定性。分布式数据库大量使用本地盘，而本地盘因为硬盘固件异常、存储介质异常、数据跳变或硬盘达到使用寿命等原因，可能引发数据库响应迟缓、数据丢失，甚至数据库“夯死”和业务中断。然而，当前许多服务器缺乏专业的亚健康检测、故障处理和自愈恢复等可靠性保障机制，进一步加剧此类故障的风险。

3.2.4 系统高可用能力存在短板

采用本地盘部署的数据库通常通过将集群分布于不同的机

房并利用逻辑日志复制的方式来实现容灾。然而，在实际应用中，网络抖动/服务器异常等因素常导致同步复制降级为异步复制，进而引发数据库集群主备两端数据不一致的问题。因此，在实际操作中，很多金融机构选择暂时关闭日志同步功能，目的是优先保证业务的顺利运行，而后再重启容灾功能，虽然保证业务的即时需求，但暂时牺牲数据库的高可用性。另外，在主从架构下，主备节点之间的主从切换可能会造成业务的短暂中断或应用报错，影响业务连续性。例如，某国有大行采用此种容灾方案时，频繁遭遇大事务堵塞、慢 SQL 执行或主备网络异常等问题，导致主备数据库数据不一致，最终只能通过业务回溯来进行数据补偿。

为应对这一挑战，业界主流做法是通过引入专业存储来实现存算分离架构的创新，助力实现多读多写、多副本归一、双集群容灾等关键技术突破，解决服务器以及硬盘的可靠性问题，如坏盘、慢盘和磨损等，同时提升资源利用率。

针对这些瓶颈，我国数据库厂商已经开始对数据库架构进行优化。例如，OceanBase 通过存算分离部署，可满足金融级高可靠要求，在存储节点接口卡、控制器故障及存储 IO 路径阻塞等场景下，数据库切换时间可达到秒级（ $RT0 \leq 8s$ ）。具体而言，数据库层的集群内主节点故障时，从节点能自动升级为主节点，主从切换时间为秒级。在链路层通过 NoF 组网实现存储与服务器间故障场景的秒级切换。在存储层采用全互联 A-A 架构，LUN 无归属，控制器之间实现负载均衡，故障倒换时间极短。在软硬件结合，服务器、存储和 NoF 网络的全面加持下，数据库性能大幅提升。

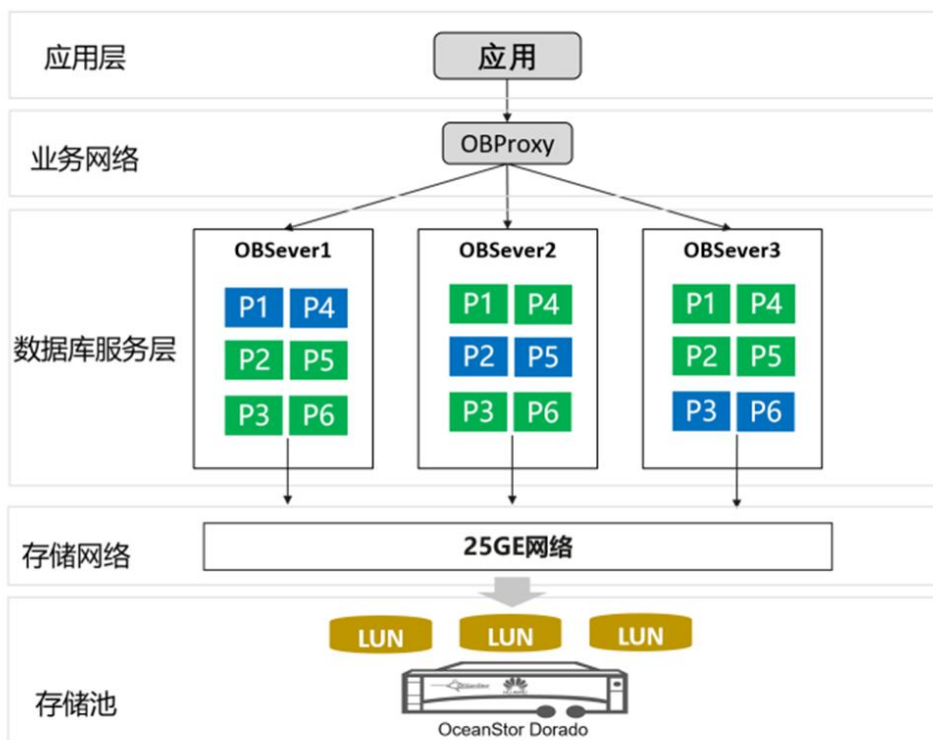


图 13 “OceanBase+专业存储” 存算分离方案架构图

华为 GaussDB 存算分离双集群方案能够实现数据库高可靠容灾，确保 RPO=0。

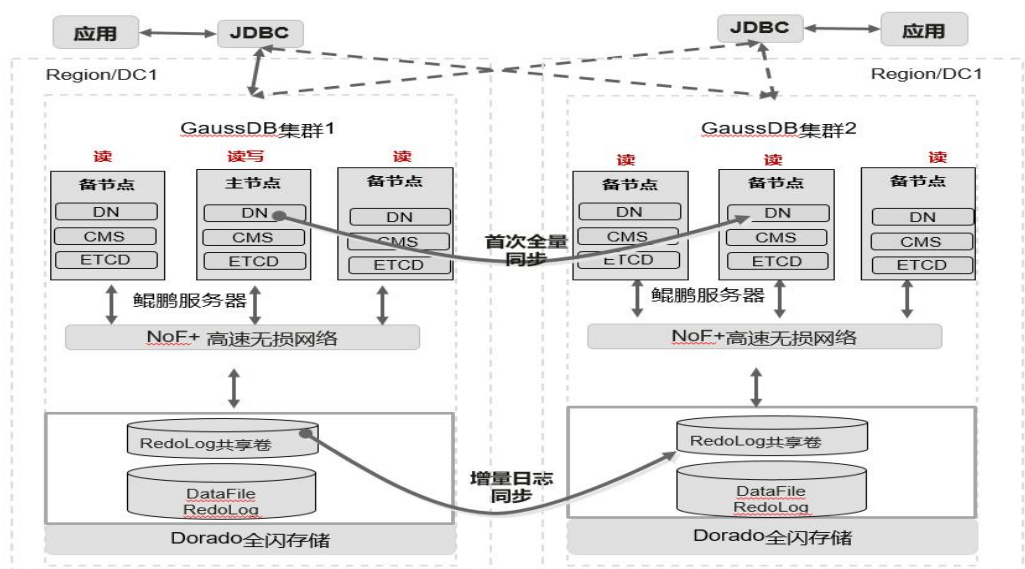


图 14 “GaussDB+专业存储” 存算分离双集群方案架构图

3.3 云原生数据库存算分离技术

云原生数据库的快速发展与其独特的架构优势密不可分，主要体现在以下几个方面：一是高效利用硬件资源，通过资源池化来提供大容量、资源利用率高且具有高性价比的服务；二是具备极致弹性，支持计算和存储等组件的独立扩展，实现无服务化；三是具备出色的数据库容灾能力，确保数据库的高可用性。



图 15 云原生数据库存算分离架构图

云原生数据库的设计理念是构建一种更契合“资源弹性管理”理念的数据库架构，充分利用云平台的资源池化特性，适应云平台基础设施，实现资源规格的灵活控制、应用的多模、更优的弹性扩展能力以及有效的成本控制。云原生数据库适用于多种应用场景，包括满足高弹性流量需求、保障数据库高可用、实现高时效性、应对混合负载挑战以及满足弹性化、智能化的成本管理需求等。其中，计算和存储分离架构是核心，而存储能力的建设是关键。

3.3.1 多级高可用，提供异地灾备能力

具备多级高可用能力是云原生的核心要求，这对于抵御不同级别的异常情况、保障客户业务的平稳运行至关重要。云原生数据库通过提供全球数据库网络服务、支持单个实例跨可用区部署、多副本、跨可用区和跨地域的多可用区热活高可用等技术，实现数据库容灾。这些技术保障了即使在主集群或任何可用区发生地域级别故障的情况下，存储一致性也不会被破坏。存储作为云原生数据库底座，其可用性和可靠性直接影响系统整体。

下图展示了 PolarDB 采用计算与存储分离的设计理念，满足公共云计算环境下根据业务发展弹性扩展集群的需求。在 PolarDB 中，数据库的计算节点仅存储元数据，而数据文件、Redo Log 等则存储于远端的存储节点。各计算节点之间仅需同步 Redo Log 相关的元数据信息，显著降低主节点和只读节点间的复制延迟。在主节点故障时，只读节点可以快速切换为主节点。

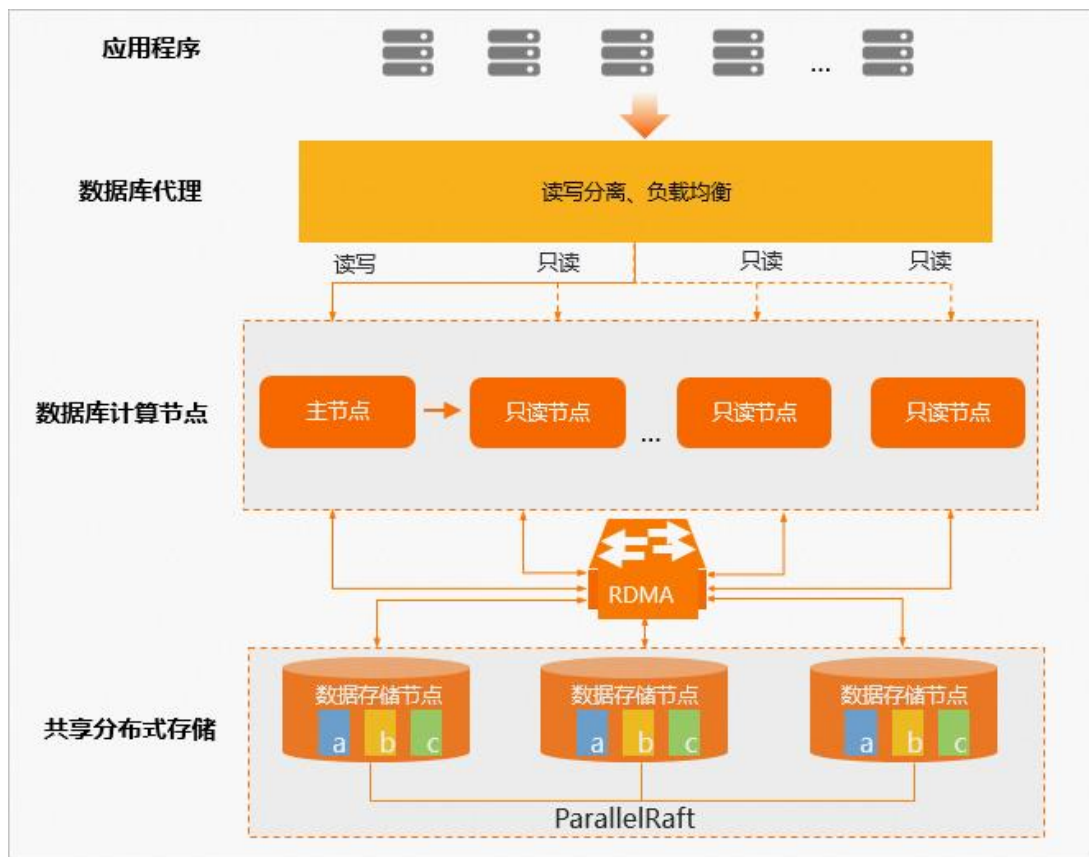


图 16 PolarDB 存算分离方案架构图

3.3.2 多级 HTAP 释放潜能

整合存储资源池、内存资源池技术与 HTAP 是云原生数据库领域的一大发展趋势，融合了 OLTP 和 OLAP 能力的技术，结合了存储池化和内存池化等软硬协同优化手段，可以显著降低网络吞吐量，提高数据库的性能和响应速度。下图展示了 GaussDB 通过存算分离部署来实现 HTAP 能力。

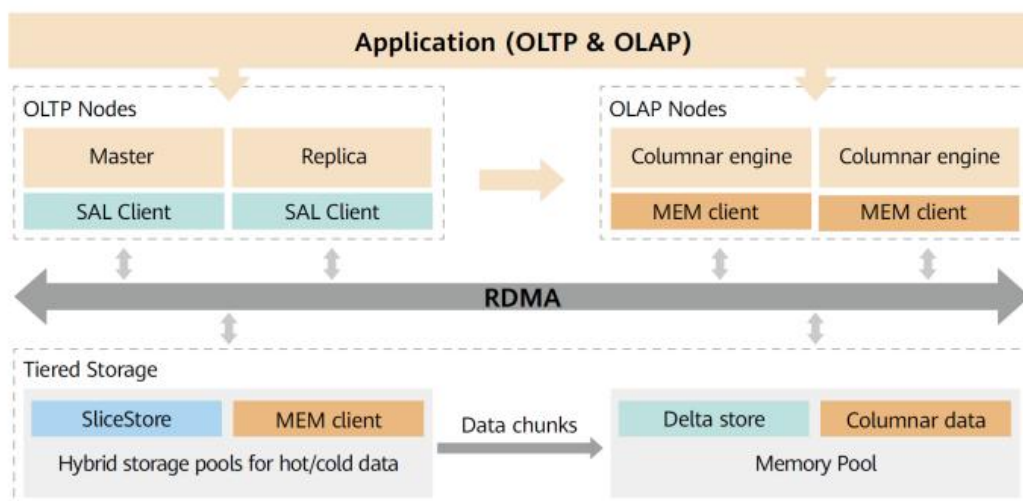


图 17 GaussDB HTAP 存算分离架构图

将存储池化和内存池化技术与 HTAP 相结合，可以为 HTAP 提供更为稳定和高效的支持，主要体现在以下几个方面。

一是以存强数，进一步提升数据库性能：使用专为数据库优化的存储结构，不仅可以消除传统存储阵列的瓶颈，还能用于卸载数据库处理工作；存储服务器中的 CPU 可用于加速数据库密集型负载；通过将算子下推至存储，使存储参与 SQL 运算，提升数据库整体性能。

二是减少网络通信开销：通过内存池技术，可以将数据缓存在本地内存中，减少对远程存储的访问和网络通信开销，提高系统整体性能和响应速度。

三是提高数据一致性和可靠性：存储池叠加内存池技术可以保证数据的一致性和可靠性，避免因网络延迟或数据同步不及时导致的数据不一致问题。

四是实现资源共享和动态扩展：通过存储池化和内存池技术，可以实现资源的共享和动态扩展，使系统可以根据实际需求灵活

扩展和收缩，提高系统的可扩展性和可用性。

腾讯 TDSQL-C 通过实施计算存储分离，构建共享存储架构，实现云原生资源的解耦和池化，进而提供资源独立弹性扩展能力。

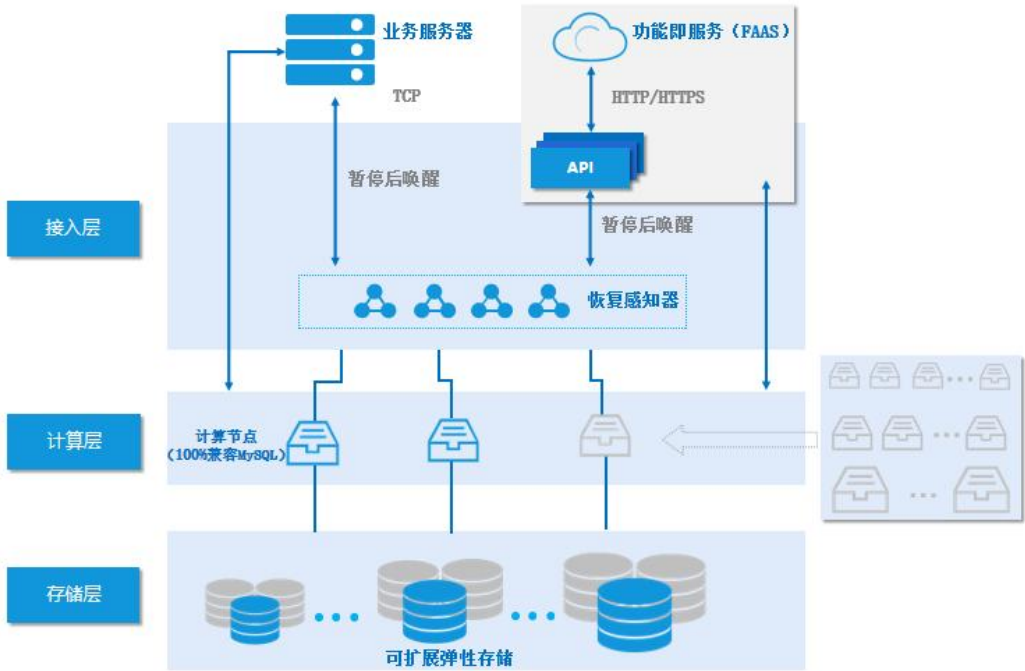


图 18 腾讯 TDSQL-C 存算分离架构图

3.3.3 智能弹性助力降本增效

随着云计算技术的不断发展，Serverless 已成为云原生数据库的重要发展趋势。然而，仅提供无服务器计算无法满足所有用户的多样化需求。因此，Serverless 数据库需具备智能弹性能力。智能弹性指数据库能够基于历史负载数据，自动构建用户画像，快速预测未来的负载曲线，并预先准备相应资源以实现弹性伸缩。使得数据库能迅速响应需求变化，防止负载触及资源上限，减少资源浪费。智能弹性是 Serverless 数据库未来的关键能力之一，有助于数据库在处理大量数据时保持高效性能，确保

系统的稳定性，降低运营成本，提高资源利用率，提升用户体验。

3.3.4 AI 助力构建智能化数据库

未来的云原生数据库将深度融合 AI 技术，通过 AI 助力云原生数据库提效，以及云原生数据库与向量能力的结合，朝着全场景智能数据库的方向迈进，为各种应用场景提供更加智能、高效和安全的数据管理服务。在这一背景下，AI 存储和具备高可用、高性能的专业存储将发挥重要作用，有效实现数据库的智能化洞察、评估和优化，确保数据库服务的安全、稳定及高效。这意味着数据库将能够自动检测和诊断问题，进行自我调优和运维，同时保证数据的安全性和可靠性。

3.3.5 云原生数据库变得更加普惠

云原生数据库正处在快速发展阶段，云服务提供商正致力于提升产品的性价比、性能和稳定性。金融机构在选择云原生数据库时，需要考虑迁移成本、技术可控性、对云厂商的依赖程度以及总体成本等因素。目前，大部分云原生数据库仍然依赖于云服务提供商自身的技术底座与硬件环境。未来，云原生数据库应朝着普惠化方向发展并具备以下特点：低门槛、开箱即用；提升迁移效率，支持跨平台、跨地域部署；与各类应用和工具良好兼容以及数据库本身的高稳定性等。

4. 数据库存算分离架构选型评估指标与建议

在进行数据库存算分离架构选型时，应当通过一系列评估指标来进行综合比较，不同业务系统对指标要求的侧重点有所不同，因此这些指标的选择应根据具体的业务需求和使用场景来确定，以确保所选架构最符合实际应用要求。

一般来说，金融行业普遍关心的技术指标有以下几类：性能、可靠性、可用性、成本收益、能耗、扩展性、安全性和运维等。

4.1 性能指标

数据库系统涉及金融行业大量的交易数据、客户信息和风险管理信息。金融机构对数据库性能指标有着极高要求，需要数据库系统不仅能处理大量的数据请求和交易，还能快速响应查询请求。关键的性能指标包括以下几个方面。

4.1.1 数据库性能

数据库响应时间：数据库对请求的响应时间，包括读取数据、写入数据、查询数据等操作的响应时间。数据库实例性能不低于 50 万 tpmC@10ms (90th)。（注：90th 表示 90%比例）。

（1）数据库吞吐量：数据库在一定时间内处理的数据量，包括读取数据、写入数据、查询数据等操作的吞吐量。

（2）数据库并发数：指同时访问数据库的用户数量。

4.1.2 网络性能

(1) 网络带宽：网络设备的传输带宽，包括数据库到计算节点的网络带宽。

(2) 网络延迟：网络设备的传输延迟，包括数据库到计算节点的网络延迟。

4.1.3 存储性能

(1) 存储响应时间：存储设备对请求的响应时间，包括读取数据、写入数据等操作的响应时间。存储可支撑部署多套数据库实例，提供稳定时延小于 0.1ms 的存储性能。

(2) 存储吞吐量：存储设备在一定时间内处理的数据量，包括读取数据、写入数据等操作的吞吐量。全读场景，提供不低于 20GB/s 的存储性能。

(3) 存储并发数：同时访问存储设备的用户数量。

4.2 可靠性指标

数据库可靠性是金融业系统稳定运行的关键。除了要保证传统数据事务的正确性和可靠性，还需要关注整系统的可靠性。数据库业务连续性和业务体验要求数据库系统在遇到各种异常和错误情况时，能够维护数据的一致性和完整性。数据库可靠性能力是指数据库系统在面对异常和错误时，确保数据的一致性和完整性的能力，目标是实现 99.9999% 的数据可靠性标准。

4.2.1 数据可靠性

(1) 数据正确性：数据库记录的信息始终保持与描述对象一致，在多个实例间不会出现数据冲突或矛盾。

(2) 数据一致性：在多个副本和不同节点之间，数据保持一致的程度。

(3) 数据完整性：数据在存储、处理和传输过程中保持其正确性和一致性。

(4) 数据持久性：存储设备故障或其他异常情况下，避免出现数据丢失，数据保持不变且可恢复。

4.2.2 系统可靠性

(1) 冗余设计：计算节点和存储节点故障不会影响整系统运行。

(2) 自动故障切换：在计算极端或存储节点发生故障时，系统自动切换到备用节点，以保障数据库服务的连续性。

(3) 网络链路冗余：链路端到端故障秒级切换，存储 IO 不归零，数据库业务不中断。

4.2.3 硬盘可靠性

(1) SSD 盘寿命提升：具备 SSD 盘磨损均衡和全局反磨损均衡能力。

(2) 最大可容忍 3 盘同时故障，3 盘故障业务不归零，性

能无明显影响。避免一个 RAID 组内 3 块 SSD 同时失效导致的数据丢失，提高系统整体硬盘的可靠性。

（3）故障预防：硬盘故障提前识别，提前告警，提醒用户进行更换。

（4）故障检测&容错：提供硬盘在线诊断能力，检测硬盘潜在故障，识别盘故障，主动启动数据拷贝。避免慢盘进行写操作，尽量不读慢盘。出现慢盘后，业务不归零，性能无明显影响。

（5）盘故障修复：由于采用存算分离架构，单盘故障数据重构速度高于 5TB/h。

4.3 可用性指标

数据库是数据业务的核心基础设施，因此必须保障其高可用能力。《信息安全技术 信息系统灾难恢复规范》对信息系统的灾难恢复目标进行了明确要求：“灾难恢复目标是通过有效的技术与管理手段，确保组织在灾难发生时迅速恢复信息系统运行，最大限度地降低损失和影响，保障数据的完整性和可用性，保障业务的连续性。”。具体来说，数据库高可用性是指在系统发生故障时，数据库仍然能够持续提供服务的能力。这对确保应用系统的稳定运行至关重要。同时，高可用能力可以减少数据丢失和业务中断的风险。为了实现数据库高可用，需要采取一系列措施，如数据中心内集群高可用容错、数据中心间集群冗余保护以及数据库备份恢复等。这些方案旨在最大限度地减少系统停机时间，

保证业务连续性和数据可靠性。

为了保证金融业信息系统的安全稳定运行，金融管理部门及相关行业协会在信息系统灾难恢复标准制定方面制订了多个管理规范 and 指引，以加强金融业信息安全保障体系建设，规范金融领域信息系统灾难恢复工作。金融业监管机构对信息系统灾难恢复标准的政策发展和变迁，经历了从要求运行安全，到制定信息系统灾难恢复预案制定、开展应急演练，再到建立完善的应急保障和业务连续管理体系的过程。

4.3.1 容错能力

(1) 数据副本：支持多副本数据存储，数据可在多个节点之间进行同步。

(2) 实时监控：系统运行状态的实时监控能力，当系统出现异常时自动报警或通知。

(3) 故障切换：计算节点发生故障后，系统自动切换到备份节点并对外提供服务。实例集群内故障切换指标：本地同城高可用方案可实现 $RPO=0$ ， $RT0<120S$ 。

4.3.2 集群冗余

(1) 同城双集群：基于共享存储的高可用方案，提供性能更高，可靠性更强的高可用方案。支持同城双集群数据库强同步，实现 $RPO=0$ ， $RT0<120S$ 。

(2) 两地三中心：支持同城双数据中心，多地容灾，实例可以在多个地域间进行迁移。实例跨地域故障切换指标，两地三中心异地高可用方案实现 $RPO < 1$ 分钟， $RT0 < 10$ 分钟。

4.3.3 备份与恢复

(1) 备份效率：支持 LanFree 级备份，数据备份流量走存储网络，降低对主机的影响，备份速率可达 2GB/s。

(2) 恢复时间：数据恢复所需的时间。支持任意时间点恢复，百 TB 级数据库分钟级恢复。

(3) 数据丢失点：数据恢复时允许的数据丢失量。支持分钟级最小保护周期，通过备份数据恢复数据库实例允许丢失的数据量最小可达到 5 分钟。

《信息安全技术信息系统灾难恢复规范》将灾难恢复能力等级划分为六个等级。数据库系统在灾难恢复能力等级需要满足第 6 级要求，即实现数据零丢失和远程集群支持。《银行业信息系统灾难恢复管理规范》在此基础上，规定了银行业信息系统灾难恢复应遵循的管理要求。具体要求如下。

表 1 银行业 $RT0/RPO$ 与灾难恢复能力等级的关系示例

灾难恢复能力等级	RT0	RPO
1	2天以上	1天至7天
2	24小时以上	1天至7天
3	12小时以上	数小时至1天
4	数小时至2天	数小时至1天
5	数分钟至2天	0至30分钟
6	数分钟	0

4.4 成本效益分析

成本效益也是选型时重要的考量因素。金融业需要在数据库的性能、安全性与成本之间寻找平衡点，选择性价比最优的数据库产品。衡量数据库存算分离架构的成本效益时，需要综合考虑性能、资源利用率和运营成本等多个方面，确保在满足业务需求的同时实现有效的成本控制。

4.4.1 成本效益指标

（1）资源利用率：存算分离架构允许资源按需分配，避免资源浪费。重点关注计算节点、内存池和存储池的利用率，以确保资源得到最佳利用。

（2）硬件成本：存算分离架构通过实现多个计算节点共享同一份存储数据，有效降低了存储成本。可以通过比较存算分离架构与传统存算耦合架构的硬件成本，评估节省的金额。

（3）维护成本：存算分离架构中，各个组件的维护工作相互独立，不再受制于相同的故障域。这虽然降低了维护成本，但也需要考虑跨节点通信的延迟对性能的影响。

（4）扩展成本：增加计算或存储资源的成本和复杂性。

成本效益测算举例说明：对比同城高可用容灾的两种方案：

1) 采用服务器本地盘架构，需要部署两套 MySQL 集群，每套集群一主两从，总共需要 6 台服务器；2) 采用存算分离共享存储架构，2 台服务器+1 套外置存储即可实现。两个方案的成本效益

测算对比如下。

表 2 成本效益计算参考

	存算一体方案			存算分离方案		
	设备数 (台)	高度 (台 /U)	高度合计 (U)	设备数 (台)	高度 (台 /U)	高度合计 (U)
计算	12	2	24	4	2	8
交换机	2	1	2	2	1	2
存储				1	4	4
硬盘框				2	2	4
合计			26			18
空间节省	-30.8%					

结果表明：在机房空间占用方面，存算分离方案明显可以减少空间占用。

4.4.2 存储空间利用率

存储空间节省：本地盘三副本和专业存储 RAIDTP（允许三块盘故障）冗余对比，以 100T 数据为例，约减少 1/2 裸容量。

表 3 空间效益计算参考

	本地盘存储（三副本）	存算分离专业存储
计算	300TB	
存储		149.3TB
合计	-50%	

4.5 能耗指标

自双碳目标提出以来，中共中央对金融业绿色低碳发展提出了明确要求。然而，采用存算一体多副本技术架构的数据库会导致数据量急剧膨胀数倍，加之服务器盘无法共享，往往需要使用小容量盘，从而致使硬盘数量、服务器数量激增，能耗也随之快

速上升。这不仅难以满足金融业绿色低碳发展要求，还会导致运营成本上升。在数据库存算分离架构中，通过资源池化、多副本归一等技术，可以减少数据副本的数量，降低容量消耗，进而实现能耗的降低。能耗指标包括：

（1）服务器能耗：服务器在运行数据库存储和计算任务时所消耗的能量。

（2）网络能耗：网络传输数据时所消耗的能量，包括服务器之间的通信和客户端与服务器之间的通信。

（3）存储设备能耗：数据库存储设备在读写数据时所消耗的能量，包括硬盘、固态硬盘等。

（4）计算设备能耗：数据库计算设备在进行数据处理时所消耗的能量，包括 CPU、GPU 等。

能耗节省测算举例说明：基于同城高可用容灾场景，对比存算一体和存算分离的两个方案的能耗指标：1）采用服务器本地盘架构，需要部署两套 MySQL 集群，每套集群一主两从，总共需要 6 台服务器；2）采用存算分离共享存储架构，2 台服务器+1 套外置存储即可实现。两个方案的能耗节省测算对比如下。

表 4 能耗节省计算参考

	典型功耗（W）	存算一体方案	存算分离方案
计算	800	9600	3200
交换机	400	800	800
存储	2295		2295
合计		10400	6295
功耗节省		-39%	

结果表明：服务器数量越多，空间节省越大。若服务器本地盘为两套 MySQL 集群，功耗预计节省 39%。

4.6 扩展性指标

金融行业的数据量非常庞大，且还会持续增长。因此，选择具有良好可扩展性的数据库显得尤为重要。在数据库存算分离架构中，扩展性是指系统在增加负载或扩大规模时，能够保持性能和功能的一致性与可靠性。

4.6.1 计算节点扩展性

(1) 节点数量：系统能够支持的计算节点数量。

(2) 线性扩展性：增加计算节点时，性能是否能够线性提升。

(3) 负载均衡：计算节点之间的负载均衡能力，确保没有节点过载或空闲。

4.6.2 存储扩展性

(1) 存储容量：系统能够支持的最大存储容量，支持按需扩展数据库存储容量，数据库业务无感知，新增容量不影响原有数据库业务性能。

(2) 存储节点扩展：系统的性能能够随着节点数量的增加而线性增长，当系统中新增节点时，系统的性能提升能够线性扩展到新增节点的 90% 以上。

4.6.3 数据一致性与高可用性

(1) 数据一致性：系统在扩展过程中，维持数据一致性和

性能的能力。

(2) 故障恢复：存储节点或计算节点故障时数据恢复能力。

(3) 数据副本：数据副本机制及其对性能和可靠性的影响。

4.6.4 资源调度和管理

(1) 资源调度策略：计算任务与存储资源的调度和分配策略可按需扩展。

(2) 自动扩展：系统是否支持自动扩展计算和存储资源。

(3) 管理工具：提供监控和管理系统性能、故障和资源使用情况的工具和接口，可预测扩展策略，确保系统始终具有良好的扩展性。

4.7 安全性要求

由于金融业的信息系统关乎国计民生，因此对数据库系统的安全性要求极高。人民银行发布的《中国人民银行业务领域数据安全管理办法（征求意见稿）》明确要求：“除因业务影响、产业制约，并可提供详细分析报告情形外，应当优先采用商用密码技术对信息系统中第三层级以上数据项实施加密存储，结构化数据项在对数据库文件整体实施加密基础上鼓励进一步采用更细粒度的加密方式，非结构化数据项可仅对拆分的第三层级以上结构化数据项单独实施加密”。原有通过应用改造进行加密的方案将对业务产生较大的性能影响，使用存算分离架构，通过外置存储的加密方案既能保证数据的有效加密又能降低对应用的性能

影响。存算分离架构将计算和存储分离，而独立的存储资源管理有助于更好地保护数据的安全性。同时，数据持久化存储的勒索防护和数据访问控制有助于减少数据泄露和攻击的风险。

（1）数据加密：对于敏感数据，需要进行加密存储，确保数据在传输和存储过程中的安全。提供硬盘级加密，性能影响小于 5%。提供存储阵列级加密，性能影响小于 20%。

（2）数据防篡改：在数据库的数据存储区域，防止攻击者对数据产生的数据文件和日志文件进行修改。

（3）侦测分析：检测 I/O 行为是否异常，如发现异常，能及时感知数据被勒索病毒加密，立即进行主动防御，防止攻击进一步扩散。100%业务负载时，侦测分析对性能影响下降小于 30%。

（4）访问控制：具备完善的访问控制策略，限制用户的访问权限，确保只有授权用户能够访问数据。

（5）身份验证：提供身份验证机制，确保用户的身份合法，防止恶意攻击和数据泄露。

（6）安全审计：需要建立安全审计机制，记录数据访问和操作的日志，确保数据的安全性和完整性。

（7）网络安全：基于零信任安全模型，建立网络安全机制，防止网络攻击和数据泄露。

（8）硬件安全：需要保证硬件设备的安全性，防止硬件故障和数据丢失。

4.8 运维指标

随着金融数据中心信息系统规模逐渐壮大，金融数据中心逐

步从最初的单数据中心延伸扩展到多地多中心运营的模式，运维工作的复杂性和繁重程度也日益增加。传统的局部、粗放、碎片化的 IT 运维管理模式难以满足新形势下业务连续性保障的实际需求。在金融业数据库架构选型过程中，考虑运维方面的相关指标非常重要。具备良好运维能力的数据库系统可以有效减轻运维部门的工作量，降低运维复杂度，为金融业的数据库系统平稳运行提供保障。

（1）独立扩展：存算分离允许独立扩展存储和计算资源，运维团队可以根据需求，动态调整资源，避免资源浪费。

（2）故障隔离：计算和存储的分离意味着单个组件的故障不会直接影响到整体系统，便于故障排查和恢复，提高系统的稳定性。

（3）灵活性：运维人员可以根据不同的工作负载优化存储和计算配置，进行精细化管理，提高资源利用率。

（4）简化维护：分离架构使得系统维护和升级更为简单，运维团队可以在不影响整体系统的情况下，对某一部分进行维护。

（5）易于实施容灾：存算分离架构更容易实现多活部署和灾备方案，提升数据安全性和业务连续性。

（6）集中管理：集中管理存储资源，提高数据管理效率，简化数据备份、恢复和迁移操作。

基于这些优势，存算分离架构能够在复杂的运维环境中提供更高的灵活性和可靠性。

4.9 演进能力

金融等行业在商业数据库应用方面已积累了数十年的经验，对商业数据库提供的稳定性能有着深厚的依赖。更换数据库产品通常涉及复杂的应用改造和业务切割，这可能带来业务风险。在当前复杂多变的国际局势下，不确定性因素激增，数据库架构选型需要考虑长期演进能力，以规避潜在的断供风险。

（1）全栈软硬件自主可控：实现软硬协同，数据库与存储深度适配，支持方案长期演进。

（2）行业部署能力：适应和满足不同业务场景和行业需求，技术架构在诸多行业得到验证。

（3）生态兼容性：能够与多种技术栈、工具、平台和服务进行无缝集成和互操作。

4.10 总结

在金融业的业务系统中，数据库承担着关键的角色，涉及大量的数据存储和处理任务，同时需要保证数据的安全性和可靠性。因此，在评估存算分离架构和存算一体架构时，需要从多个维度进行深入对比，以全面理解两种架构的特性。

表 5 数据库存算分离架构 vs. 传统架构选型指标对比

选型指标	存算分离架构	存算一体架构
性能	高性能，存储和计算资源独立优化。由于计算和存储可以独立扩展，性能瓶颈可以更容易地被缓解。	性能有限，受制于单一节点的能力。计算和存储资源紧密耦合，在某些情况下可能会出现资源争用问题。

可靠性	存储和计算分离，提高系统的故障恢复能力。通过独立的存储节点，数据可以更可靠地存储，且存储和计算节点的故障不会相互影响。	故障恢复较慢，单一节点故障影响大。虽然整体架构较简单，但计算和存储在同一节点上，某一节点的故障可能会影响更多服务。
可用性	高可用性，通过冗余和分布式架构提高系统可用性，计算和存储节点的独立性使得故障隔离更容易。	由于耦合度高，节点故障的影响范围较大，可用性受限，单一节点的故障可能导致系统不可用。
成本效益	初期投资较大，但长期运营成本较低，更灵活，可节省资源。	初期投资较低，但扩展和升级时成本可能较高。
能耗	初期能耗较高，但长期运行能耗较低，由于资源利用率高和优化的资源分配。	初期能耗较低，但长期能耗较高，因为资源利用率较低，容易产生闲置。
扩展性	高扩展性，计算和存储资源可以独立扩展。	扩展存在一定限制，需要同时扩展计算和存储资源。
安全性	更强的安全性控制，可以分离安全域，独立控制存储和计算的安全策略，安全策略可以更细粒度地实施。	整体安全策略较易实施，安全性较低，单一节点的安全漏洞影响整个系统。
运维	按需动态调整资源，简化维护，集中管理，便于实施容灾方案。	管理复杂度高，故障处理成本高，容灾方案实施难度高。
演进能力	适应快速技术变化，独立升级计算和存储模块。	演进能力有限，升级需要更换整个节点。

总的来说，存算一体架构在业务发展初期，适合中小规模业务，可满足快速上线。但在业务大规模上线后，存算分离架构在性能、可靠性、可用性、成本收益、能耗、扩展性、安全性和运维能力等指标上具有比较明显的优势。因此，金融业需要根据具体的业务发展阶段、应用规模和运行环境进行权衡和选择，以确保选用的数据库架构能够最佳地满足金融业的需要，实现可持续稳健发展。

5. 数据库存算分离架构演进方案

实施存算分离架构要充分考虑业务的持续平滑演进，并着眼

于高可靠、高性能、绿色节能等长远目标。对于具备条件的场景，优先采用共享存储的技术，避免数据迁移和二次改造。根据当前数据库能力，演进方案可分为两种，分别适用于不同的业务目标。

5.1 外置存储的存算分离架构

通过外置存储和计算资源的分离部署，结合存储的高可靠特性能力，提升整体业务的可靠性。

分离数据库进程和数据库数据，将数据库进程部署在我国服务器中，数据库数据则通过外置高可靠高性能存储，结合高速 NoF+网络，构建统一的数据库存储资源池。支持多种类型的数据库接入。同时，利用基于 RoCE 技术的高性能网络，提升数据传输能力，实现数据库 IOPS 性能提升，IO 时延下降。

适用场景：适用于当前采用存算一体架构的数据库，通过此方案可以实现可靠性增强，并灵活进行算力和存储的独立扩容。



图 19 外置存储的存算分离架构

具体实施步骤：一是通过存算解耦实现资源池化，根据业务实际需求配置资源；计算层支持多节点横向扩展，按需采用裸金属或虚拟化部署，最大限度提升资源利用效率；二是采用容器化、虚拟化技术实现计算节点的资源隔离。通过企业级存储，保障数据中心内及数据中心间的数据一致性和完整性；三是采用标准 SAN 协议对接企业级集中式存储，将数据高可用保障下沉到存储层，计算与存储之间通过高速网络互连以保障高性能。四是架构支持主流数据库类型，以及 ARM、X86 等多计算平台。

5.2 存算协同的存算分离架构

在采用外置存储方案基础上，通过数据库和外置存储能力更加紧密地结合，软硬协同，以提升容灾能力、运维能力、成本优

化以及性能。

场景 1：客户关键类业务系统，两地三中心部署，要求同城 RPO=0，同城双集群部署，故障隔离，可以采用数据库和存储协同，实现数据库日志通过存储复制跨中心强同步，实现数据零丢失。

场景 2：现网的大库集群，容量在 10 至 60TB 左右，采用数据库层备份，全备窗口超过 12 至 24 小时，无法满足客户备份窗口要求。客户希望采用 Lan-Free 方式备份，提升备份效率。通过数据库和存储协同，数据库提供快照备份接口，做备份时间戳和数据库事务标记，存储通过快照读取备份数据，备份性能可达 2GB/s，10 倍提升备份效率。

场景 3：客户批量改造现网数据库，若采用 share nothing 数据库，存储容量成倍增长，节点和容量压力大，希望集群支持副本归一，降低整网建设成本。通过数据库和存储协同，数据库实现集群内多节点共享存储，存储实现数据本身的高可靠高性能。

场景 4：数据库全栈改造升级，需要在满足当前业务性能基础上，也要支持未来 3 至 5 年的业务增长。通过数据库和存储、计算、网络的全栈协同，实现计算、存储池化，数据库实现多读多写架构，达成极致性能和可扩展性。



图 20 存算协同的存算分离架构

6. 金融业数据库存算分离架构应用实践

6.1 工商银行数据库存算分离架构实践

6.1.1 发展现状

工商银行金融产品数量多，服务客户体量大，对应系统数量庞大和体系复杂，服务连续性要求高，且广泛应用与数据库产品耦合度较高的存储过程等高级特性，历史资产重。涉及总行应用接近 200 个涵盖核心业务、渠道、前置、外联、管理与支撑等多类业务系统，其中高等级应用超过 90 个。

6.1.2 面临挑战

传统数据库有其极强的功能黏性、优秀的系统稳定性、良好的软硬适配能力和对金融业务的良好支撑，且金融业存量系统往往与数据库特性高度耦合，大量业务逻辑内嵌到数据库实现，且

具有历史比较久远、业务长期稳定、关联应用较多等特点，在数据库转型过程中面临诸多挑战。

（1）服务连续性能力要求高

金融核心应用 7*24 小时对外服务不中断，传统数据库及国际主流的通用基础应用产品，具备优秀的系统稳定性和可靠性，能支持高性能、大容量业务的长期稳定运行。而自主创新软硬件产品尚处于快速迭代、加速成熟的过程中，在保证金融业务服务连续性方面，对我国数据库的数据可靠性高、系统可用性、集群性能及服务连续性能力提出更高要求，需要保证数据库转型过程中的生产安全和迁移后长期稳定可靠运行。

（2）高级特性和存储过程重度使用

传统数据库转型替代难度较大，尚无成熟的原位替代解决方案，金融机构多采取对应用系统进行分布式改造，并对业务实现逻辑进行重构的方式实现转型。工商银行合计使用超过 40 类传统集中式数据库对象、24 类基础数据类型、200 个系统内置函数、70 个系统高级包和 200 个系统视图，且广泛适用自定义类型、自定义函数、递归、嵌套、多表关联、自连接等众多高级语法特性，存储过程代码总行数达到亿级规模，数据库对象总数量超千万。数据库转型过程中，面临技术复杂度高、工作量大、项目周期长、实施风险高等挑战。

（3）业务逻辑复杂

金融业存量业务系统具有历史比较久远的特点，经过长期演

进，业务逻辑复杂，且与其他业务系统的关联关系复杂，交易调用链路相互交织。业务系统数据库转型往往伴随着存量技术资产、业务资产、关联关系梳理，涉及数据库对象迁移改造、应用代码改造、新旧系统双轨并行并叠加业务系统之间交易链路对接等工作，甚至涉及应用架构和业务逻辑重构等工作，技术复杂度高。涉及海量数据库对象数量、存储过程代码行数、应用程序代码行数和广泛使用高级语法特性，要保证迁移后系统功能对等、性能满足要求，应用迁移改造工作量大。自主创新软硬件产品的性能、稳定性、成熟度等方面与传统商用软硬件产品相比还存在一定的差距，确保转型的平滑、稳定、安全，成为数据库架构转型中的难点，需做好业务双轨并行、灰度分批切流和应急回切等措施来规避。

6.1.3 转型实践

（1）突破数据库关键技术

1. 构建分布式数据库精简模式部署形态，实现稳定性性能规格突破

工商银行对标主机 DB2 和 Oracle 数据库，按照“分布式架构+集中式体验”的理念，基于全栈自主创新软硬件产品，实现分布式数据库的技术突破。

一是软硬件协同，提升性能。通过多核并行、基于代价的分布式并发事务并发优化和全链路并行编译执行等，实现了分布式

查询技术优化，良好地兼顾了金融业务系统高并发联机业务场景和复杂计算批量场景的需求，具备百万级 QPS 能力。

二是采用存算分离架构打破单库容量瓶颈，实现弹性扩容。通过计算存储分离、存储资源池化、在线弹性扩展等，具备 PB 级海量存储能力，良好地支撑了历史明细类、报表等金融业务专题的海量数据处理需求。

三是构建精简模式支持集中式数据库原位替换。构建精简模式（分布式数据库的集中部署），优化集中式部署性能可靠性，满足业务对部署形态的不同需求，支撑大量存量业务原位替换需求，同时相同产品多部署形态实现开发、运维能力的统一构建。

2. 提升容灾能力构建多级容灾体系

为确保应用转型能力服务能力、安全稳定性能力持平，容灾能力有提升，工商银行对标原主机 AB 站点双活方案，基于存算分离架构与厂商进行大型业务系统 Oracle 数据库转型的多集群方案技术攻关，首创建成多集群多中心部署的高可用架构，满足金融业务系统服务连续性要求。

一是同城双园区间通过存储级复制实现增量数据强一致同步，异地园区间通过异步方式实现增量日志同步，具备本地 $RPO=0 \& RT0 < 30$ 秒、同城 $RPO=0 \& RT0 < 180$ 秒和异地 $RPO < 1$ 分钟 $\& RT0 < 10$ 分钟的金融级高可用通力。

二是实现双集群故障和运维隔离、支持灰度发布。基于同城双园区双集群部署架构，实现双集群间跨数据库版本的相互兼容

和增量数据的强一致同步,具备业务不中断情况下数据库版本灰度升级通力,在数据库版本升级异常情况下,具备 RPO=0&RTO<180秒的回切能力。同城双集群署架构示意如图所示。

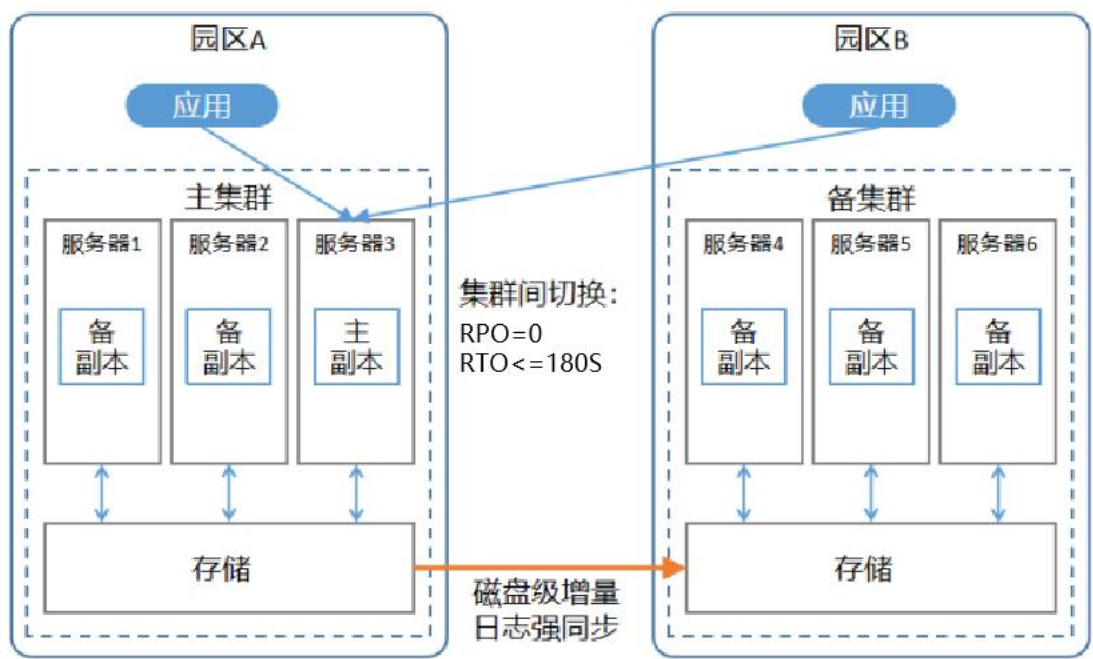


图 21 同城双集群署架构示意图

目前工商银行综合应用存储级日志复制、数据库流复制等多种技术,形成多地多中心多集群容灾方案体系,满足不同等级金融业务系统的容灾需求。分级容灾方案示意如图所示。

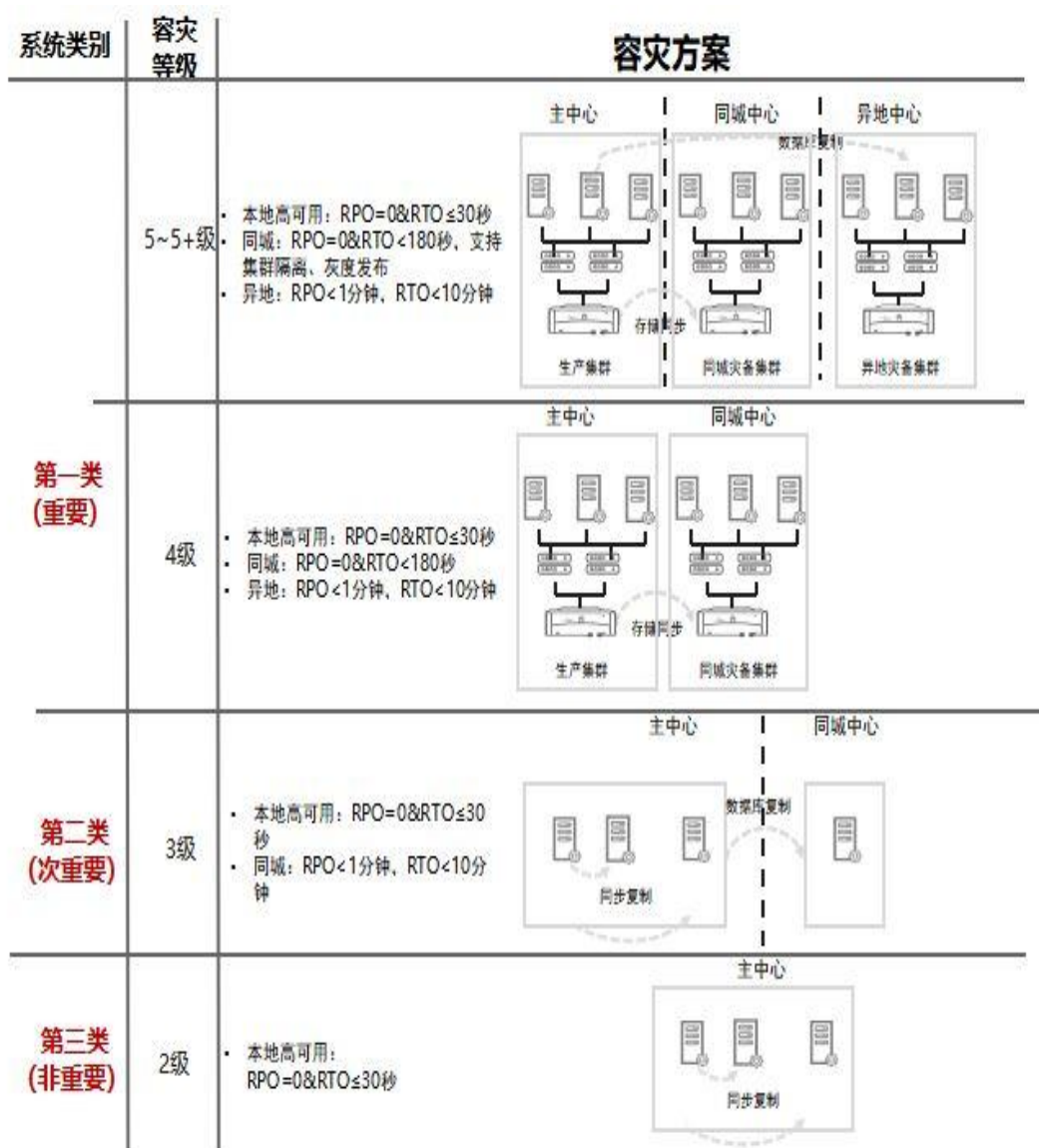


图 22 分级容灾方案示意图

(2) 构建原位替代模式降低存量迁移难度

针对存量业务累积构建大量业务逻辑内嵌到数据库内分布式改造难、存储过程迁移难等问题，工商银行采用稳态业务原位替换模式有效降低技术复杂度和转型工作量。

1. 稳态业务原位替换，平滑迁移

对有存量存储过程的业务，优先采用精简模式，以集中式部署形态进行平移替换，对于当前部分国产数据库并发性能无法满

足的巨石类业务，优先考虑进行数据库拆分，保留存储过程，从架构层面避免或减轻了迁移改造的诸多挑战。

降低改造难度，减少工作量：避免了大量存储过程、复杂 SQL 改造，减少应用控制分布式事务等方面的负担，同时由于是架构平替，降低了存量业务工具化迁移难度，进一步减少工作量，使有限的精力可更聚焦于业务。

降低应用迁移风险：降低了历史大量累积的业务逻辑在迁移时出现问题的风险。在业务逻辑保持不变的情况下，可以通过自动化测试工具、流量回放工具减轻测试工作量，更好保障迁移质量。

2. 敏态业务分布式改造

对业务增长迅速，可通过分片较理想实现并发、容量提升，且无存量存储过程、复杂查询的业务，可进行分布式改造，采用分布式数据库替换。

（3）实现平滑迁移

为解决迁移转换工作量大，测试覆盖难等问题，工商银行联合头部科技企业开展 Oracle 数据库平滑迁移技术攻关，通过多试多用，不断总结经验，研发全流程自动化迁移、自动化测试、自动化仿真验证、灰度切流工具，配套建设完备的技术资产社区和全流程标准化转型工艺，实现了复杂数据库特性和巨量存储过程的自动化迁移能力，人工改造成本压降 90%以上，有效解放了生产力，让科技力量更加聚集于金融业务创新和数字化转型领域。

自动化迁移：实现数据库表、索引、存储过程、序列、视图、触发器、自定义类型等数据库对象和高级包函数、自治事务、递归调用、自连接等复杂特性的自动化迁移能力，以及全量和增量数据的自动化同步，自动化迁移成功率达 95%以上，人工改造成本压降 90%以上，突破了数据库转型的技术瓶颈和实施障碍。

自动化测试：研发覆盖单元测试、功能测试、性能测试、生产验证和测试管理过程的自动化测试工具链，降低测试人力投入和测试复杂度，提升测试效率，程序测试覆盖率达 100%，分支覆盖率达 95%，保障数据库架构转型过程平稳可控。

自动化仿真验证：构建交易录放工具，通过一致性流量回放和性能回放，仿真阶段实现业务功能全覆盖测试和接近实际生产业务压力的性能、可用性及可靠性测试。

灰度切流：建设异构数据库数据复制工具，实现异构数据库间存量数据和增量数据双向复制。在双轨运行阶段，通过业务增量归档数据在异构数据库间的双向复制，实现新旧系统业务数据的准实时一致，确保故障场景下能及时回切，提升对外服务的连续性。数据同步效率可达 150GB/小时。

6.1.4 转型效果

经过大量实践，工商银行形成了一套无需整体重构 Oracle 存储过程逻辑，低成本、高效可控的原位替换转型技术方案、配套工具和转型方法论，构建全金融业务场景支撑能力，广泛用于

包括个人网银、信贷系统、贵金属等 130 多个业务系统，覆盖办公系统、一般业务系统和关键业务系统各类业务专题。采用精简模式、存算分离等架构设计简化复杂问题，通过数据库与国产众核处理器、企业存储协同构建的高性能、高可用、高安全、弹性伸缩、双集群容灾能力以及自动化数据迁移、同步工具能力数据库产品商业版本中落地，具备快速推广的能力。

6.2 农业银行新一代存算分离数据库实践

6.2.1 发展背景

在金融数字化转型浪潮中，银行业务对信息系统的各项处理能力都有较高要求。这些系统不仅承载着农业银行的核心业务，如账务处理、支付结算和信贷管理等，也确保了内部管理系统的高效运作。数据库作为金融信息系统的“记忆中枢”，负责存储和检索海量金融数据，处于技术架构的核心，是连接前端应用和后端基础资源的关键纽带。

数据库按存算功能是否解耦可分为两种，一种是计算和存储耦合的存算一体数据库，优点是计算在本节点读写数据，数据传输速度快，缺点是不易灵活扩容。另一种是计算和存储解耦的存算分离数据库，优点是能够实现计算资源与存储资源的独立配置和扩展，缺点是存储和计算之间大量数据的传递对 IO 要求较高。此外，存算分离架构与云基础设施技术结合，使得数据库的管理与运维能够充分利用云计算的优势，实现资源的自动化动态弹性

伸缩和快速按需分配。

6.2.2 实践经验

基于存算分离数据库的技术特点，农业银行结合自身需求特点与实际情况，采用策略规划、产品验证、行内落地这一思路整体推进数据库的应用落地，聚焦如下关键方面。

一是制定分类的硬件资源配置策略。结合存算分离技术特点与行内的资源供给及基础设施情况，针对不同使用场景制定不同的使用策略。部署方式上看，虚拟机云服务资源利用率较高，物理机裸金属性能强大。存储方式上看，分布式存储部署灵活，企业级高端存储的性能强大，产品稳定性、高可用性更好。农业银行对于不同灾备等级的系统，设计了不同的架构。高等级系统更注重系统的业务连续性，选择“物理机裸金属+企业高端闪存”的组合架构。低等级系统在稳定运行的前提下更注重资源利用率，采用“云服务虚拟机+分布式存储”的组合架构。

二是完成产品的全面验证。农业银行基于数据库的存算分离架构，进行了详尽的功能与性能测试，旨在提前识别并迅速解决潜在的产品问题。与存算一体数据库系统相比，本次测试的范围更广泛，内容更复杂，和传统测试具有较大差异性。功能测试方面，除了验证数据库的传统基本功能外，还结合云计算技术，增加了对资源动态管理能力的测试，以确保系统能够灵活配置资源。对于高可用，考虑到计算和存储的解耦，增加了计算节点切换机

制、存储部分可靠性与冗余性的测试。对于数据备份恢复，存储分离技术带来了更灵活和复杂的要求，为此增加了计算与存储节点的数据一致性校验。性能测试方面，不仅涵盖常规的性能评估，考虑到计算和存储之间的分离特性，引入存储网络性能测试，以检验在高并发、大数据量场景下的存算分离数据传输效率与稳定性。此外，还特别注重资源扩展与回收的性能测试，旨在验证系统在面对动态资源需求变化时的响应速度与效率，确保业务连续性和用户体验。

三是推动我国数据库在行内的落地。进行产品验证的同时，完成存算分离数据库与我行内部运维管理体系的融合工作，确保实现运维无缝集成和高效协同。在提升产品稳定性方面，为预防存储设备的单点故障风险，引入存储池化和日志卷双写等特性。为进一步提升产品易用性，引入集群单浮动 IP 和跨集群单浮动 IP。在提升运维效能方面，农业银行对标行业顶尖实践，在完成基础运维体系搭建的基础上，积极探索并实施自动化运维策略与工具创新。按照脚本化、自动化、智能化三步走的策略，构建智慧运维体系，实现了运维流程的全链路监控与分析。在精细运维方面，农业银行充分利用云计算的弹性扩展优势，结合持续迭代的高阶运维功能，对数据库容量实施了标准化的动态管理，为业务的长期稳健发展提供了坚实的支撑。

6.2.3 实践成果

经过一年多的努力，农业银行基于存算分离数据库，完成涵盖信贷管理、资金转账、金融市场交易等多个关键领域各类不同灾备级别系统的投产运行，覆盖京、沪、蒙三地多中心。自投产以后系统运行稳定，各项关键性能指标与功能需求均满足各类业务和技术要求，标志着农业银行基于存算分离架构的数据库使用方面取得了重大进展。

展望未来，农业银行规划将存算分离数据库技术的应用范畴进一步拓展至快捷支付、超级柜台服务、电子银行等关键业务领域，旨在为全行提供更加稳定、高效、智能的技术保障，驱动银行业务实现更加稳健与快速的发展。

6.3 平安银行 MySQL 数据库存算分离架构探索

平安银行当前主流架构为存算一体，数据库使用开源 MySQL，通过 vmware 虚拟化进行资源隔离，操作系统是 REDHAT，存储使用的是本地盘，主备同步使用 MySQL 数据复制技术，建设的标准是两地三中心一主四备的架构。

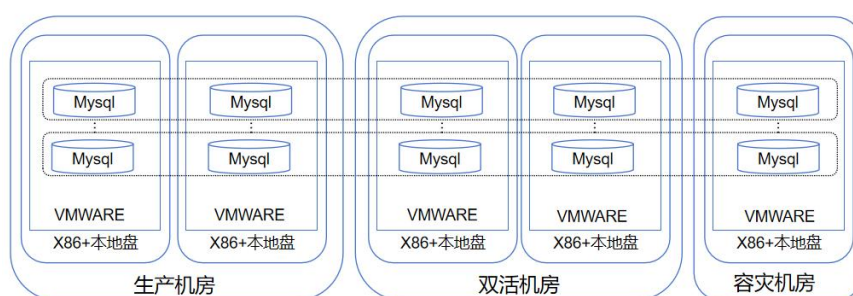


图 23 两地三中心一主四备的架构示意图

6.3.1 面临痛点

一资源使用率不均，由于虚拟化后单物理机上部署多台虚拟机，但是每个虚拟机上运行的 MySQL 数据库因业务而异，导致的计算资源与存储资源分配和使用不均衡，整机的资源使用率经常会出现计算或者磁盘先到达瓶颈，而剩下的资源却比较空闲的情况，导致整体的资源使用率难以提升。

二是故障场景下 RPO 难以满足，因为传统的 MySQL 架构基于数据库事务日志的传输和回放，在故障场景下的主从切换，因为日志的同步存在网络时延、大事务等客观影响因素，而且在繁忙的数据库体现尤为明显，切换的 RTO/RPO 可能会到 1 分钟到 15 分钟甚至更久，难以满足金融业的需求。

三是受制于国外商用软硬件，虽然 MySQL 为开源数据库产品，但是使用的 VMWARE 虚拟化技术、红帽操作系统、服务器硬件都是国外商用软硬件，在基础架构层技术仍然受制于人。

四是虚拟机 OS 过多，如果选择我国 OS 会显著增加成本。

6.3.2 解决方案

方案一：存算分离虚拟化 MYSQL

这种架构是对当前 MySQL 架构的改进，但是需要从下而上解决之前的问题。一是设备采用 ARM/海光信创服务器；二是采用

我国虚拟化技术；三是不再使用本地盘、改用外接的我国存储；四是采用开源或者我国操作系统。

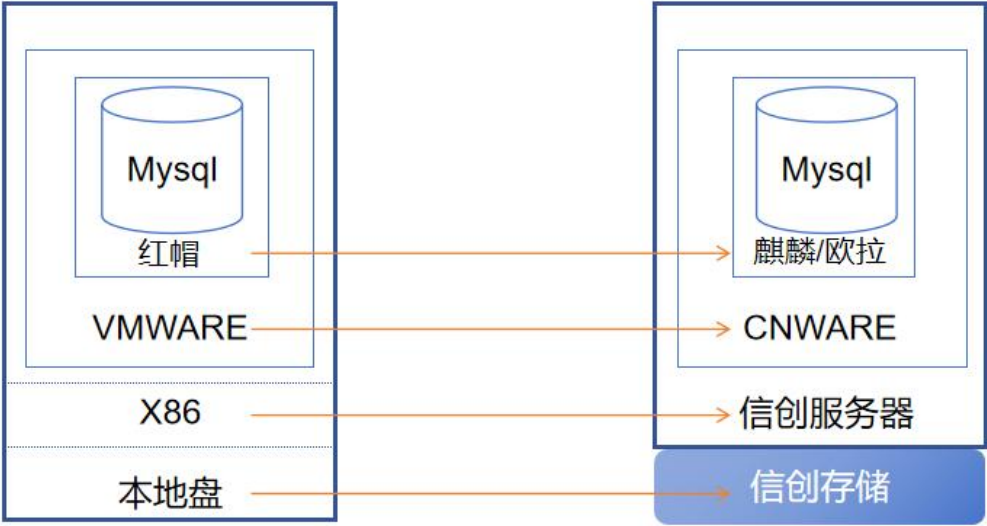


图 24 升级改造路径示意图

上图是规划的建设路径，路径上各个环境的方案已经测试验证通过，存算解耦只能帮助解决计算资源和存储资源不均衡的问题，但是无法解决数据一致性的问题。

解决数据一致性问题，目前有两个思路。一是通过数据库层的强同步实现，MySQL 最多只能实现半同步复制，无法实现强同步复制，所以需要使用数据一致性保障能力更强的数据库版本。二是通过存储层的数据共享，实现主写备读的模式，来实现主库故障情况下保障数据一致性，实现 RPO=0。以上两种方案各有优缺点，还在进一步研究和验证过程中。

针对资源使用率的问题，如何避免过高和过低，还在研究如何合理调度通过削峰填谷来实现资源有效利用，可能会建设一套智能化的资源管理系统。

方案二：容器化 MYSQL

MYSQL 上容方案，可能是未来的部署方案，目前还只是在测试方案探索阶段，有很多问题需要解决。设想通过容器化实现数据库服务化，数据库可按需进行资源动态的扩缩容管理，实现 Serverless 能力。

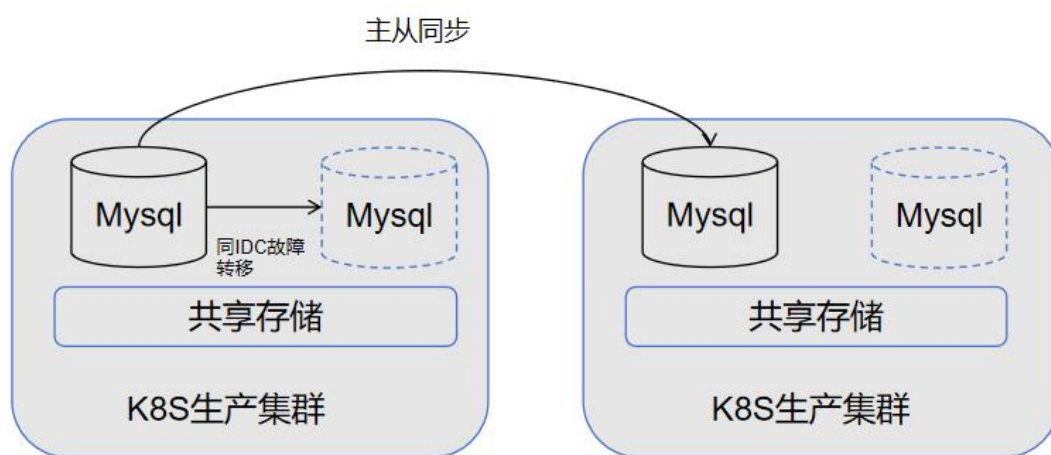


图 25 容器化后架构示意图

上图为容器化后 MYSQL 的架构方案，该方案使用外接共享存储，在上容的同时使用存算分离架构，是一种更加彻底的存算分离。该架构下同 IDC 内不再建设有备库，主库故障后利用 K8S 的调度和共享存储能力，在原存储上可以迅速拉起新节点实现故障转移，而在同城跨 IDC 机房建设有另外一套独立的 K8S 集群，通过数据库复制技术进行跨 IDC 主从同步。该架构有以下优点：

一是 RT0 和 RP0 的提升。存算分离后，在容器化部署中，当出现主机故障时可以在原有外接存储上利用 K8S 的调度能力迅

速拉起一个新 POD 节点，由于使用同样的存储，相比传统架构可以做到数据零丢失。而在同城双活中，MYSQL 运行在另外一套的 K8S 集群和共享存储，使用数据库自身的主从复制进行数据同步。极端场景下，单 IDC 的机房、K8S 集群、主机、存储故障，都可以做到切换到双活的 MYSQL 进行对外提供服务。

二是资源利用率提升。计算资源和存储资源解耦，资源调度更具有弹性，提高资源利用的同时，可增加单台物理机的部署密度，进一步提升资源利用率。同时由于利用了存储和容器的能力，MYSQL 的从库也较传统的一主四从架构减少到一主两从，进一步减少资源的消耗。而且在容器化部署中，对于资源的弹性伸缩调度，相比传统架构会变得更加灵活。

6.4 上海银行数据库存算分离架构的应用试点实践

6.4.1 项目背景

上海银行在推进新建系统和存量系统改造过程中，基于本行应用的数据库体量，确定如下选型策略：优先集中式数据库，降低开发成本与运维复杂度；单库达到较大交易量和数据量，将分布式数据库纳入考虑。

本次试点的跨应用数据检核系统，业务层面进一步深化数据检核体系，从原先单一应用、单一维度的数据检核向多层次、多维度交叉领域拓展。实现相关金融基础数据和行内经营管理信息平台的数据交叉比对，提高数据的一体化程度和数据口径的一致

性。跨应用数据检核系统主要是数据分析，属于 OLAP 类系统，为此试点使用 TDSQL 数据库（使用集中式实例模式）和集中式存储阵列。

6.4.2 架构设计

(1) 资源池及架构设计

服务器/存储	数量	说明	共享说明
管控服务器	3	主 IDC: 3	多套系统共用一套管控
数据库服务器	2	主 IDC: 2	
集中式存储阵列	2	不同节点需要不同的集中式存储	可共享

部署拓扑:

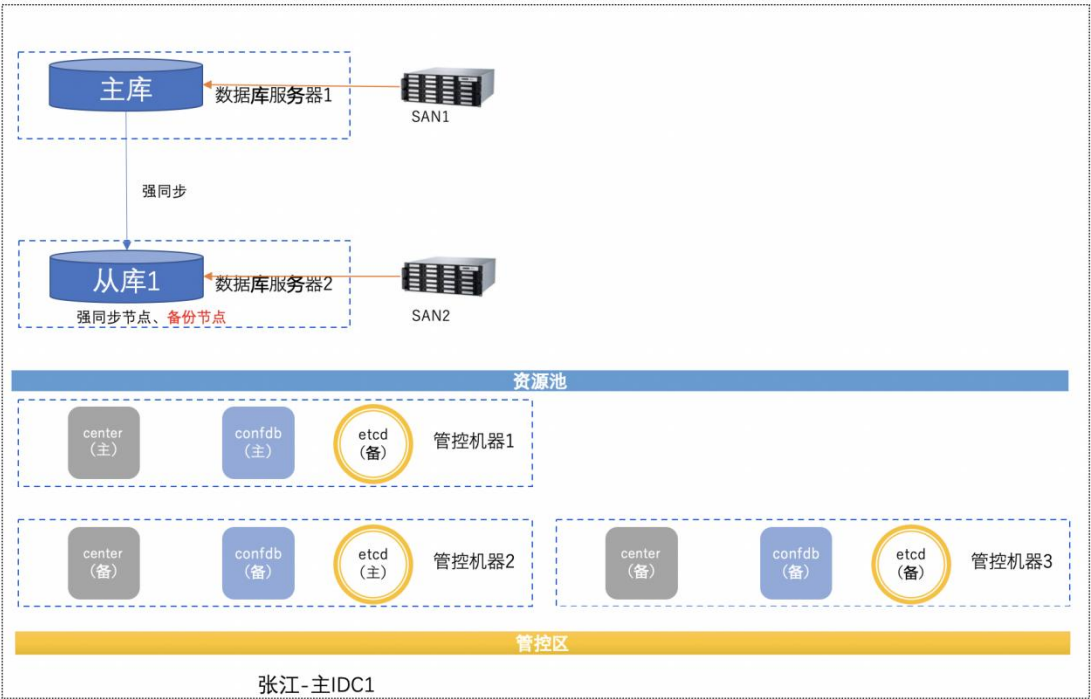


图 26 部署架构示意图

架构说明:

数据节点: 运行数据库实例，实现主从复制。

管控系统：负责 TDSQL 集群的监控运维，包括高可用切换、备份、监控、告警等功能。

(2) 存算架构选择

通过实际性能测试比较，使用集中存储阵列 (OceanStor Dorado 18500 V6)、全闪 SSD 硬盘 (NVME 协议)，相比本地 SSD 硬盘，综合性能有较明显提升 (包括 IOPS、带宽和时延等指标)，具体数据如下：



图 27 数据对比示意图

结合性能影响、容量扩展便利性、磁盘设备可靠性等因素，集中式数据库决定使用集中式存储。

6.4.3 实现收益

(1) 实现方式

存力和算力剥离，各司其职。计算节点 DN 节点负责执行应用请求、生成执行计划、数据文件读取。使用外置存储将数据进行妥善、安全的存储，即应用产生的实际数据文件存放在全闪集中式存储上。同时全链路采用轻量级 NVMe 协议连接计算节点和存储，达到充分利用 SSD 的高速读写能力和低延迟特性。

（2）效果和收益

在扩展性方面，存与算不受资源配比影响，扩展性得到提升，计算与存储按需各自独立扩展。采用集中式阵列可以提供更大的容量空间，在线使用容量扩展更为便捷。

从主机到交换机网络到存储，采用端到端 NVMe 协议通信，支持更高的队列深度和并行性，在提高存储设备的性能同时也可实现应用多底层多盘并发读写，在多线程情况下读写表现高于本地盘。

在性能方面，大数据量 IO 密集型数据库，使用集中式存储相比本地盘，性能有明显提升。在可靠性方面，通过使用可靠性更高的专用存储设备提升数据库可靠性，也降低主备切换的概率，提高可用性。在数据安全方面，单盘或多盘故障对上层不感知，数据重构不影响性能，降低 IO 的敏感度。

6.5 微众银行 TDSQL + OceanDisk 存算分离架构

6.5.1 项目背景

微众银行 TDSQL 数据库前期采用存算一体架构，在线业务数据存放于数据库服务器本地 SSD 盘，该架构存在以下几个问题。

（1）本地 SSD 磁盘空间有限，随着数据规模增长，数据库容量瓶颈较难解决；

（2）受限于本地 SSD 盘容量和 IOPS 能力，单台数据库服务器部署的实例数较少，服务器 CPU 资源相对空闲，存在资源浪费；

(3) 慢查询等大 IO 操作易导致磁盘高负载，引起数据库性能抖动；

(4) 本地 SSD 故障需进行数据库服务器替换，对业务有一定影响，每年因磁盘故障发起的自动或人工切换可达 20 次以上。

经过产品调研和 POC 测试，选用华为分布式存储设备 OceanDisk，存放 TDSQL 在线业务数据，实现 TDSQL 存算分离架构。该架构可提升 TDSQL 可靠性，提高资源利用率。

6.5.2 总体方案

(1) 部署架构

TDSQL 采用存算分离架构，计算层采用 X86 或 ARM 架构服务器，存储层采用分布式存储 OceanDisk，数据库服务器通过高速 RoCE 网络协议直接访问 OceanDisk。TDSQL 以资源池的形式交付，一个资源池包含多台计算服务器和 3 台 OceanDisk，这部分硬件资源均匀分布在 3 个不同的数据中心。生产集群的 TDSQL 实例采用一主两备的三副本架构，主节点可读写，备节点只读。主备间通过 binlog 做数据同步，由 TDSQL 内置强同步机制保证主备数据一致性。

单台服务器部署多个 TDSQL 实例。通过 cgroup 技术为每个实例设置 CPU 资源限制。每个实例使用独立的逻辑卷，借助 OceanDisk 的 SmartQoS 特性，设置单个逻辑卷的 IOPS 和带宽上限，从而保障关键实例获得优先的 IOPS、带宽和时延保障。

限制单台 OceanDisk 部署的 TDSQL 实例数量（如 50 个），降低 OceanDisk 故障影响范围。避免 OceanDisk 负载过高，保证实例获得的 IOPS、吞吐量和 IO 时延在可接受范围内。将数据库主备节点均匀分布在不同的 OceanDisk 上，实现存储负载均衡。

为确保 OceanDisk 稳定可靠运行，需控制其负载在一定阈值下。OceanDisk 采用双控制器架构，控制器 CPU 性能是常见瓶颈，日常 CPU 利用率应控制在 50% 以下。当其中一个控制器发生故障，另一个控制器可以接管其请求，保障业务稳定运行。

（2）容量管理

1. 磁盘容量规划

OceanDisk 基于全闪存 SSD 构建，共 36 块盘（36*7.68T），配置 RAID 6。应基于历史数据和未来增长预期，评估 TDSQL 容量需求。OceanDisk 整体空间使用率控制在 70% 以下，预留空间应对突增的数据量以及磁盘故障导致的可用容量下降。TDSQL 单实例挂载的逻辑卷可在线扩容，扩容上限为 6T，需避免因单实例过大导致影响范围大、恢复时间长等问题。

2. IOPS 资源管理

制定分层级的 IOPS 资源分配策略。新上线的 TDSQL 实例初始阶段设置一个较大的 IOPS 上限值，充分利用 OceanDisk 的性能优势。稳定运行一段时间后，根据实例的历史性能监控数据，分析其 IOPS 使用模式，并结合业务 SLA 要求，合理评估所需的 IOPS 能力。通过 OceanDisk 的 SmartQoS 功能，为不同 TDSQL 实

例或数据库配置 IOPS 限额，避免资源竞争，保障关键业务的性能需求。

(3) 运维管理

1. 监控告警机制

OceanDisk 内置监控告警系统，捕捉异常生成告警信息后，会上报到行内统一告警平台。

通过 Restful API 拉取 OceanDisk 的 IOPS、吞吐量、延迟、容量使用率、I/O 错误率等关键指标信息写入数据库，与 TDSQL 监控数据做联合分析：

- 设置指标告警阈值（如 CPU、IOPS 、IO 时延等），超过阈值则判定为异常；

- 捕捉 IO、CPU 等监控指标阈值趋势变化，关联导致占用资源消耗较大的 SQL；

- 采集全量 SQL 语句，跟踪 SQL 耗时变化趋势，有助于提前发现隐患 SQL。

2. 自动化运维

TDSQL 运维管理平台覆盖了 80%以上的日常运维操作，包括服务器资源管理、服务器故障替换、实例主备切换、扩缩容、节点重建、故障分析等。WEB 页面操作通过设置多重检查规则，高危操作重复确认等，有效降低运维操作风险，提升操作效率。

OceanDisk 内置 web 管理平台 DeviceManager，操作功能完善，包含了存储空间管理、任务配置、故障管理、性能管理等。

通过开发系列工具，对接 TDSQL OSS 接口及 OceanDisk Restful API，将 TDSQL 和 OceanDisk 操作连接起来，如 TDSQL 一键部署功能，可自动完成服务器和 OceanDisk 资源分配，部署全程无需干预。

6.5.3 难点问题

(1) 评估 IO 时延对数据库性能影响

OceanDisk 提供高 IOPS 和吞吐量，满足高并发业务对性能的需求。但对比本地 SSD 盘，OceanDisk 较高的访问时延会带来性能影响，如 SQL 耗时增长、数据库的主备复制性能下降。

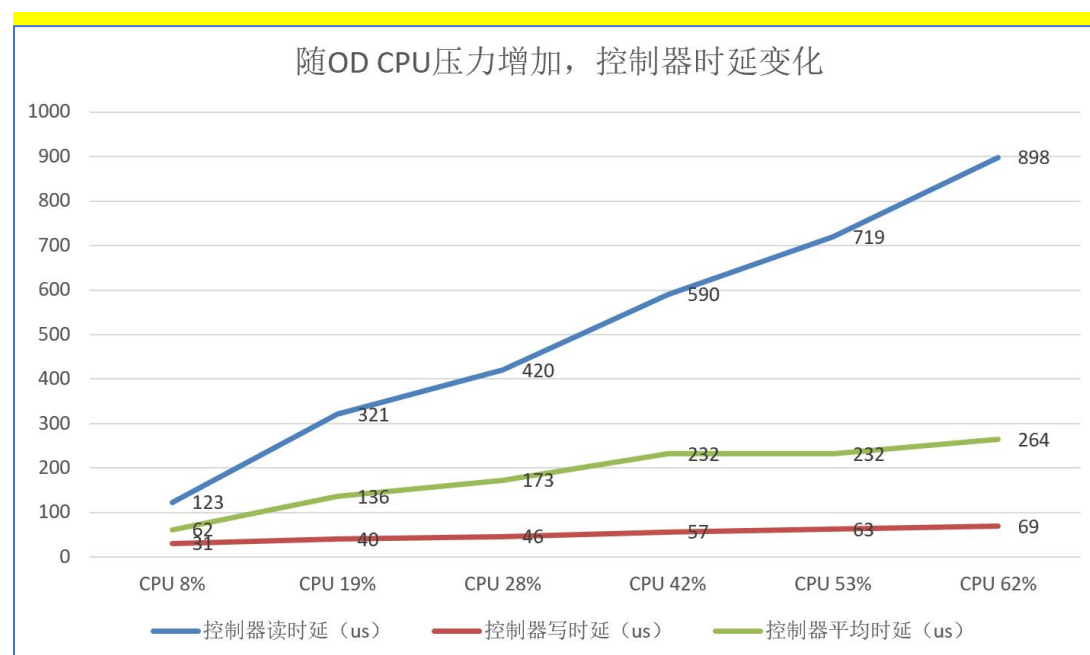


图 28 时延对数据库性能示意图

如图所示，当 OceanDisk 接收的请求压力变大，伴随着控制器 CPU 负载的增高，访问 OceanDisk 的 IO 时延也会升高，尤其体现在读 IO 读时延上。影响数据库性能，SQL 执行耗时上涨，

具体涨幅因 SQL 而异。在请求并发保持不变的条件下，随 IO 时延上升，数据库 QPS 呈下降趋势。

业务应用系统使用 TDSQL 存算分离架构前，应针对各业务场景进行全面的压力测试，分析 SQL 耗时是否在可接受范围内，以及如何优化应用数据处理逻辑，从而减少 IOPS 消耗、缓解 IO 时延增高对业务的影响。

（2）容量管理与成本控制

均衡资源使用率，提高资源利用率。1 个 TDSQL 资源组包含 3 台 OceanDisk 和多台数据库服务器。TDSQL 资源均衡包括如下 3 个层面。

- 多个资源组之间的负载均衡；

- 单个资源组内，不同服务器间的负载均衡，不同 OceanDisk 间的负载均衡；

- 单台服务器上 CPU、内存的均衡使用，单台 OceanDisk 容量、IOPS、IO 时延的均衡。

存量 TDSQL 实例资源平衡。单个 TDSQL 资源组内，收集该资源组各 TDSQL 实例的运行数据。通过分析历史资源消耗数据，计算出各实例所需资源配置。依据这部分配置数据，在 TDSQL 资源组内进行合理的 DB 节点迁移调度，使资源池内各硬件负载尽量均衡。同时应兼顾实例重要程度，保障关键实例的 IOPS 能力及 IO 时延。

新实例部署资源分配。根据新实例所需的容量、业务 TPS 等数据，通过自研算法，兼顾业务系统风险隔离以及资源均衡，选择合适 TDSQL 资源组下的 3 台服务器进行实例部署。新实例设置较高的资源上限，待运行稳定后，再评估实例实际所需资源。根据业务重要性决定如何实施资源限制。

（3）运维复杂度增加

存算分离架构下，运维需要同时关注 TDSQL 和 OceanDisk，更多监控信息，包括运行状态、性能指标和容量使用情况等。由于 TDSQL 的性能和故障问题可能与数据库自身、OceanDisk 或网络连接相关，问题排查更加复杂。

6.5.4 应用成效

（1）缓解容量瓶颈

TDSQL 存算分离架构带来更灵活的容量管理，单实例容量规格支持从 10G 到 6T，支持在线扩容。灵活容量管理也带来了空间利用率的提升。

（2）提升服务器资源利用率

借助于 OceanDisk 逻辑卷管理功能，包括 I/O 流量控制、数据压缩等。当前单数据库服务器部署实例数增加，提升了数据库服务器 CPU 和内存利用率。

（3）解决运维痛点

服务器故障无需进行 DB 节点重建，将存放数据和日志的逻辑卷挂载到新的服务器上，可直接启动新实例。缓解了因节点重建拷贝数据带来的 IO、带宽资源消耗，以及重建时间较长带来的高可用风险。

（4）TDSQL 故障率下降

OceanDisk 内置双控制器高可用架构，以及 RAID 6 配置。最大允许 2 块盘同时故障，大幅降低磁盘故障对数据库稳定性的影响。

（5）TDSQL 资源成本下降

存算分离架构增加了存储产品 OceanDisk，但大幅减少数据库服务器数量及机架机位数量。在提升数据库可靠性的同时，降低资源成本。

6.6 人保寿险数据库存算分离架构转型升级

6.6.1 项目背景

随着核心业务规模的不断扩大，面临着日益增长的数据处理需求与数据存储挑战。公司核心业务系统数据库基础设施架构主要基于传统的小型机+SAN 网络+集中式存储解决方案。然而现有传统 IT 基础设施逐渐显露出其局限性，包括设备老旧、性能表现不足、扩展性差、设备以及运维成本高昂等问题。

为了进一步提升业务处理能力，优化资源利用效率，降低运维复杂度，公司对数据库的基础设施开展全面升级改造工作。本

次改造的核心目标是将现有的小型机+SAN 网络+集中式存储的传统数据库架构，调整为基于存算分离的架构。

6.6.2 架构方案

(1) 总体架构

本项目通过选用基于存算分离架构的平台为 Oracle 数据库构建高性能基础设施环境，通过内部高带宽、低时延 RDMA 网络将计算和存储资源集成互联，实现数据的快速处理和存储，从而实现多个业务系统数据库的整合部署，提供高效、稳定的数据库服务。

目前基于该架构已部署 2 套数据库生产环境与 1 套灾备环境，涉及计算节点达 12 台，存储节点 17 台，其中单台计算、存储节点可以提供不少于 20GB/s 的吞吐。每一套环境根据容量需求使用 6 台或 5 台配置 NVMe SSD 盘的 x86 服务器部署 SmartStor 商业存储软件构建多副本全闪存储资源池，为计算层提供安全、稳定、高效的存储服务能力。本次构建的平台可以根据实际的计算、存储资源需求动态调整存储、计算资源。

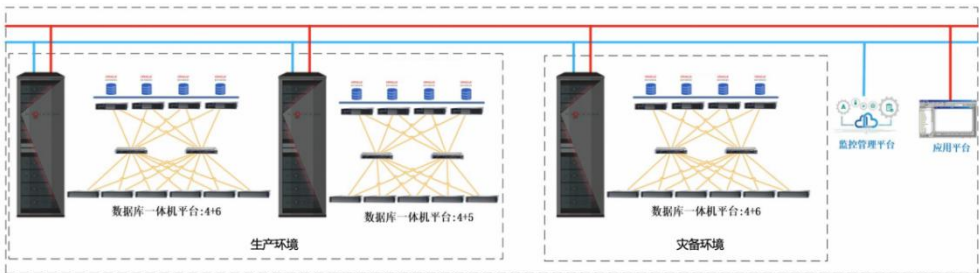


图 29 整体建设拓扑示意图

（2）功能特性

本方案架构主要包括分布式存储资源池、网络资源池、计算资源池、智能监控管理系统。各资源池提供能力概述如下：

1. 高性能存储资源池

通过构建分布式存储资源池为计算层提供存储服务能力，在架构层面具有灵活敏捷、精细化管理、资源隔离、高性能等特点；在数据保护层面通过多副本机制充分保障数据可靠性；在资源扩容层面支持在线横向/纵向灵活扩容。这些特点使得分布式存储资源池能够为企业提供高效、稳定、可扩展的数据存储解决方案，支持企业业务的快速发展和灵活变化。

2. 高性能计算存储互联网络

计算存储之间采用 RDMA 互联网络，实现了内存到内存的直接数据传输，极大地降低了数据传输的延迟。该网络支持高达 100Gbps 的带宽，具有极低的网络延迟，这使得在分布式存储环境中，数据可以在多个节点之间快速传输，可以满足各种复杂业务场景的需求并提高整个系统的性能。

3. 高性能计算资源池

计算层将数据库部署在物理服务器上，通过 RAC 机制实现数据库高可用集群，充分保障数据库运行性能和可靠性。

4. 智能监控管理平台

智能监控软件 SmartMon 对数据库、主机、存储、网络进行统一监控管理，包含数据库全生命周期管理、数据库性能监控、

数据库资源配置、数据库运维管理、数据库自动化巡检、数据库补丁升级、资源池管理、用户/权限管理、告警配置、操作审计、计算/存储/网络设备监控管理等一体化功能。

6.6.3 运行效果

根据生产系统运维数据显示，销售人员管理系统迁移至该存算分离架构后，数据库的繁忙度从 45% 下降到 5%，CPU 使用率从 70% 下降到 10% 以下；其他系统迁移完成后，该数据库的繁忙度从 50% 下降到 1%，CPU 使用率从 20% 下降到 5% 以下，单笔处理的响应时间从平均 70 余毫秒下降至平均 40 余毫秒。

根据极限压测数据显示，在 2 台计算节点，3 台存储节点的存算分离架构下，随机读 IOPS 可达 500W+，随机写 IOPS 可达 480W+，顺序读/写吞吐能力可达 40GB/s，压测 TPMC 可以达到 150 万。随着节点数的增加，其性能值接近于线性增长。

6.6.4 工作价值

采用基于 X86 架构的存算分离架构在提升系统性能的同时、通过集中的管理平台简化运维复杂度，在降低总体设备成本的同时（相对于集中式 SAN 存储），为公司的数字化转型和快速发展提供了强有力的支撑。通过此方案选型，实际效益如下。

资源高效整合与利用：通过存算分离方案，将计算资源、存储资源和网络资源解耦。分别对计算以及存储资源按需扩容，实

现了资源的优化配置和高效利用。这种整合不仅减少物理设备的数量，还通过内置的负载均衡、在线自动扩展等技术，确保了资源按需分配，避免资源闲置或过度使用的情况，从而提高整体资源利用率。

简化运维复杂度：传统的小型机+SAN 网络+集中式存储架构涉及多个独立的组件和系统，运维人员需要掌握多种技术和工具，增加了运维的复杂性和难度。而存算分离方案通过集中、丰富的管理界面，简化了系统的配置、监控、维护和升级过程，降低了对运维人员的技能要求，提高了运维效率，降低了人为导致的风险。

提升系统性能与稳定性：该存算分离架构采用优化的硬件设计，如高速缓存、专用网络等，以及针对数据库优化的软件堆栈，能够显著提升系统的处理能力和响应速度。同时，内置的高可用性和容灾机制，如自动故障转移、数据冗余存储等，确保了系统的高稳定性和可靠性，即使在面对硬件故障或灾难性事件时，也能保证业务的连续运行。

加速业务响应速度：通过提升系统性能和稳定性，存算分离数据库能够更快地处理大量的业务数据，加速业务处理流程，提高客户满意度和忠诚度。此外，它还支持实时数据分析和决策支持，帮助企业快速响应市场变化，抓住商业机会。

降低总体拥有成本：存算分离的架构，通过集中的管理平台，可以带来运维成本降低、资源利用效率提升、业务收益增加等效

益，将显著降低企业的总体拥有成本。通过减少物理设备的数量和能耗，存算分离还有助于企业实现绿色、低碳的运营目标。

7. 总结

随着金融业持续创新发展，存算一体数据库架构逐渐暴露出灵活性和效率不足的问题。为此，越来越多的金融机构开始探索存算分离架构的应用实践，以提高系统性能和可靠性。同时，数据库企业也推出存算分离的数据库产品。金融机构在进行数据库架构选型时，需综合考虑业务需求、数据量、查询复杂性、实时性、安全性和成本等因素，以选择适合自身的数据库产品。

新技术和架构的涌现显著提升数据库性能和可用性，计算、内存和存储的解耦增强系统可维护性和可扩展性，资源池化和弹性伸缩实现动态资源调整，提高利用率并降低成本。此外，独立故障域的设计提升系统可用性和容错性，减少单点故障的影响。

总之，当前数据库架构的发展趋势在于充分利用新硬件技术和高速互联网络，实现计算、内存和存储的解耦，从而提高系统的可用性、性能、扩展能力和资源利用效率。未来，存储和数据库技术将更加紧密地结合，对我国数据库的发展产生深远影响。