ArchSummit
全球架构师峰会 2017

主办方 Geekbang. 极客邦科技 InfoQ

打造金融级分布式数据库服务

钱煜明
中兴通讯 首席架构师

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

主办方 Geekbang. 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

SPEAKER
INTRODUCE

钱煜明

中兴通讯 首席架构师

钱煜明， 中兴通讯首席架构师， 20年研发管理经验，曾历任北方电讯(加拿大)，IBM(加拿大），摩根斯坦利（美国）架构师及阿里巴巴平台技术总监, 兼任中科院客座研究员, 东南大学, 南京邮电大学, 南京理工大学客座教授. 对于企业架构, 投资银行技术架构, 互联网金融, 电商平台, 电信平台, 云计算, 大数据, 分布式数据库，移动互联网,SOA等均有丰富经验. 现带领团队研发中兴通讯Golden系列数据服务平台，包括GoldenData 大数据服务， GoldenDB分布式数据库等，已在国内多家银行及金融机构部署上线。

TABLE OF

CONTENTS 大纲

- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：分布式业务连续性
- 挑战3：分布式数据安全性
- 挑战4：服务可扩展性
- 金融行业应用实践

金融信息科技“十三五”规划指明发展方向

总体目标

全面推进金融业信息化发展提质增效，积极面对新兴技术带来的机遇与挑战，主动开展架构转型，建立开放、弹性、高效、安全的新一代银行系统，深化信息科技治理成效，完善科技研发运维体系，强化信息安全和风险管理，以信息科技引领创新发展、转型升级，提升支持实体经济发展能力，促进普惠金融大发展，为社会和公众提供更加丰富、安全和便捷的金融服务。

信息科技治理有效性明显提升

- 构建信息科技治理机制评价和持续改进机制
- 组织、制度、流程和人力资源建设机制
- 科技与业务建立成熟的伙伴关系
- 优化数据治理、提升数据服务、发挥数据价值

信息科技服务能力持续提升

- 构建绿色数据中心
- 推进运维自动化和智能化
- 实现基础架构转型升级
- 建立适应互联网业务场景的软件开发过程体系

科技成为引领创新的关键引擎

- 科技创新纳入总体战略
- 大力推进“互联网+”、大数据等国家战略落地
- 构建银行业互联网金融生态
- 建立银行间和跨行业的联合创新

网络和信息安全管控能力显著增强

- 积极落实《国家网络空间安全战略》
- 实现关键基础设施基本安全、风险可控
- 健全客户信息保护机制
- 建立全方位的安全态势感知和防范应对机制
- 建立银行间网络安全协同防护机制

信息科技风险管理“三道防线”协同水平持续提高

- 明晰并落实“三道防线”的协同机制
- 构建开发、运维、安全、风险管理一体化运营平台
- 建立信息科技内部控制成熟度评价机制
- 完善外包风险管理体系
- 夯实业务连续性管理

分布式金融架构的理想世界

海量数据
平滑扩展

风险管控
数据安全

服务弹性
应对冲击

高效运维
自动容错

业务连续
快速部署

开放金融
服务社会

业务安全
数据容灾

降低成本
投资复用

数据服务架构的挑战

对数据库的高并发，可扩展性要求

- 互联网金融服务带来高的瞬时数据库并发负载，往往要达到万次/秒的读写请求，要求数据服务能够线性扩展性能与容量

数据安全性

- 数据服务需要具备安全防护手段，防止非法访问，篡改

数据服务业务连续性

- 多数据中心部署后，数据需要在灾难情况下不丢数据，保持业务连续服务

数据一致性要求

- 受监管限制，金融行业需求与互联网应用有差别，数据很多场景要求实时强一致

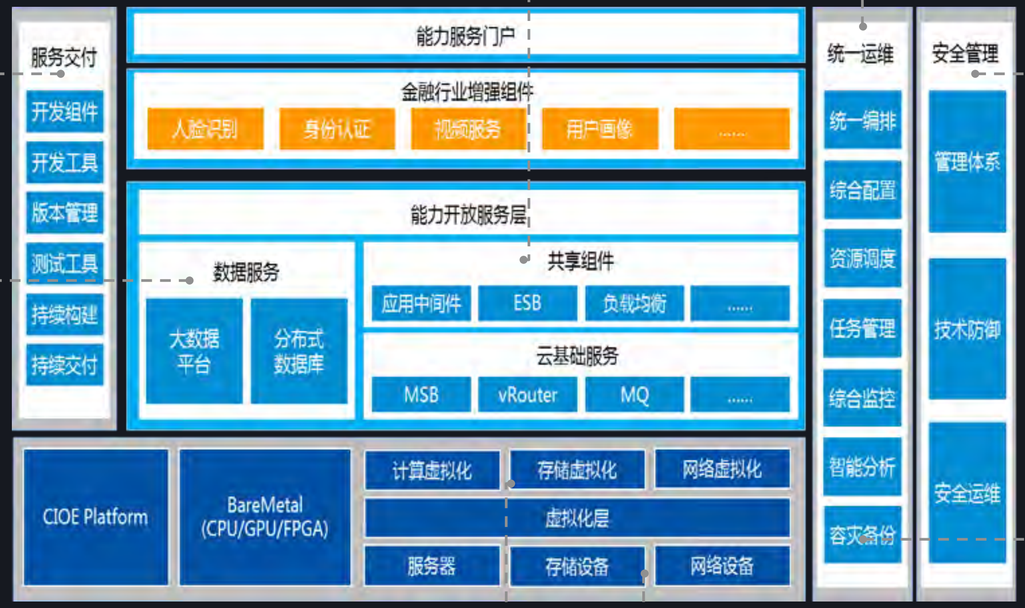
中兴通讯金融分布式架构全景视图

PaaS解决方案

DevOps解决方案

大数据平台及分布式数据库解决方案

IaaS解决方案



统一运维解决方案

安全解决方案

容灾和备份解决方案

微模块数据中心/IT硬件解决方案

面向金融的中兴通讯分布式数据库--GoldenDB

分布式数据库包括计算节点（前置中间件）、数据节点（DB）、全局事务管理器和管理节点四种组件，其中蓝色背景的为有状态的组件，需要考虑容灾：

1、计算节点（前置中间件）：提供SQL解析、优化、路由、结果汇聚、分布式事务控制等功能；

2、数据节点（DB）：真正存储业务数据的组件，通过分库分表实现数据库能力的水平扩展；

3、全局事务管理器（GTM）：分布式事务管理的重要组件；

4、管理节点（OMM与MDS）：包括元数据管理、参数配置、其他三种组件的监控与管理等。

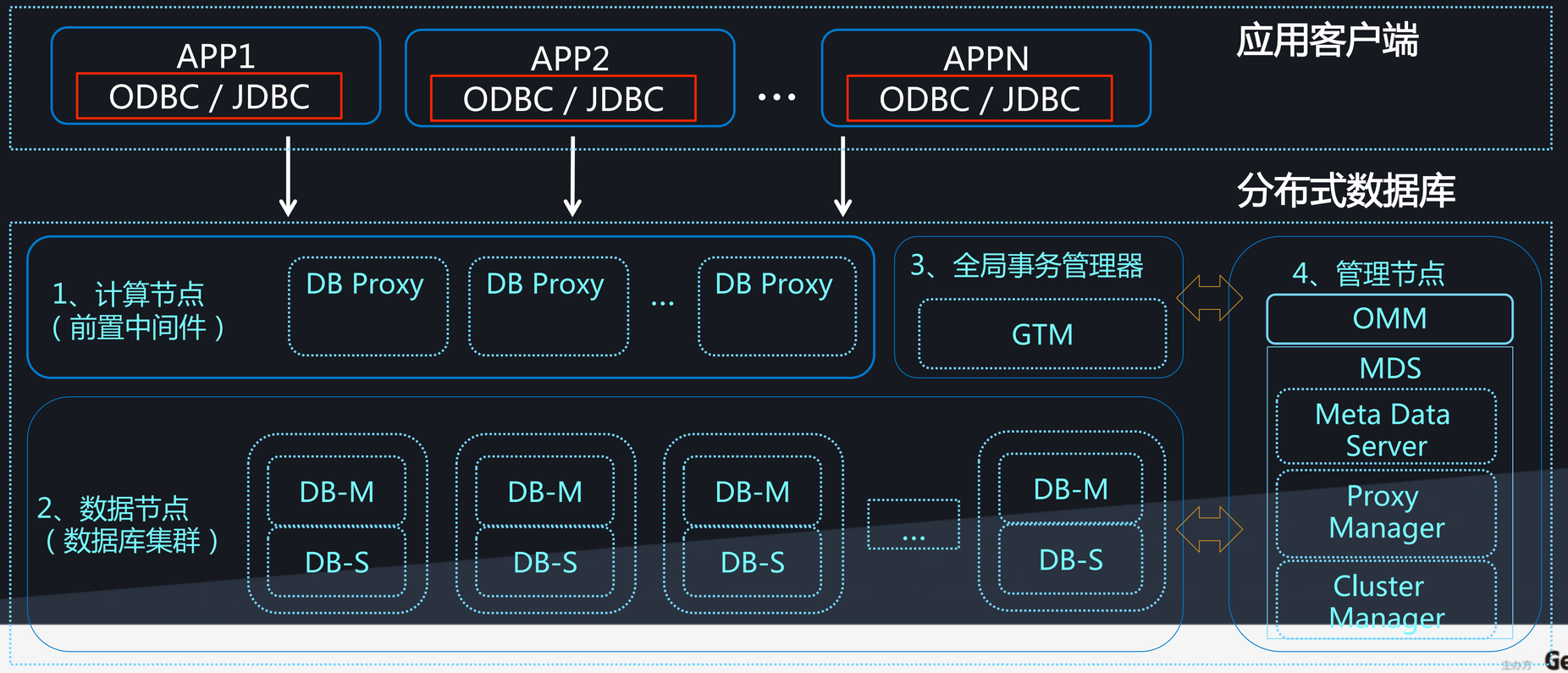


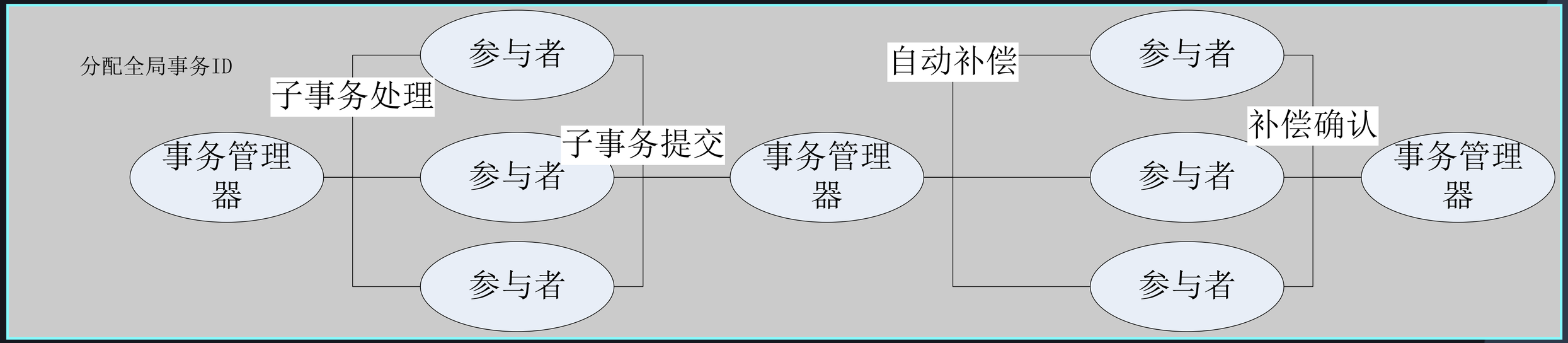
TABLE OF

CONTENTS 大纲

- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：业务连续性
- 挑战3：数据安全性
- 挑战4：服务可扩展性
- 金融行业应用实践

改进的分布式事务一致性--Sagar 模型

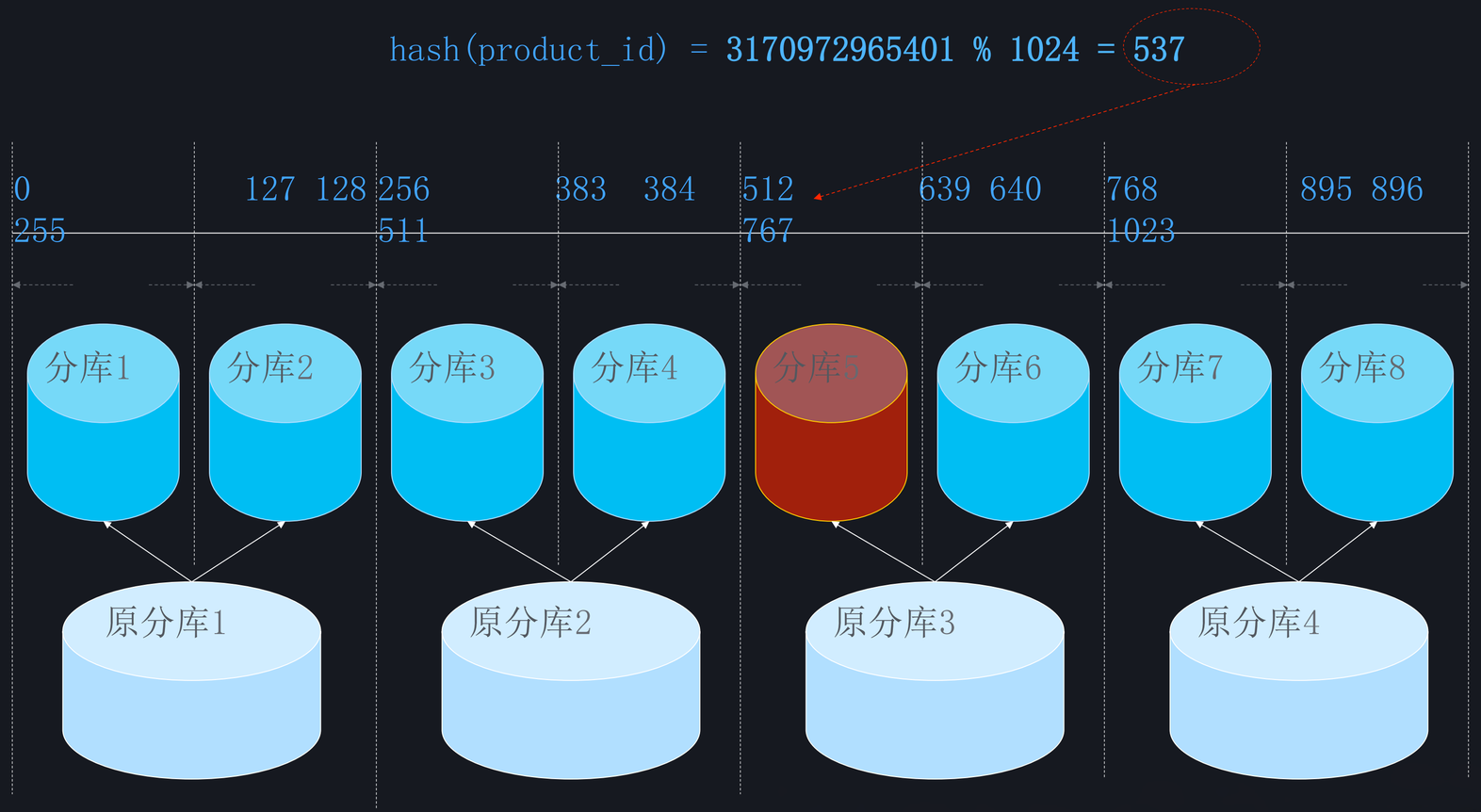
假设条件： 业务中正常情况只有极小比例事务会失败



分布式MVCC+GTM+乐观提交，完成分布式事务的全生命周期管理

- 保证各种异常场景下的数据一致性，彻底解决分布式数据库可能的脏读问题
- 自动构建事务回滚机制，从数据库日志中抓取对应的修改内容，并自动化生成补偿操作
- 在数据备份/恢复过程中，保证各全局事务组的操作统一执行或回滚

数据拆分/重分布过程一致性

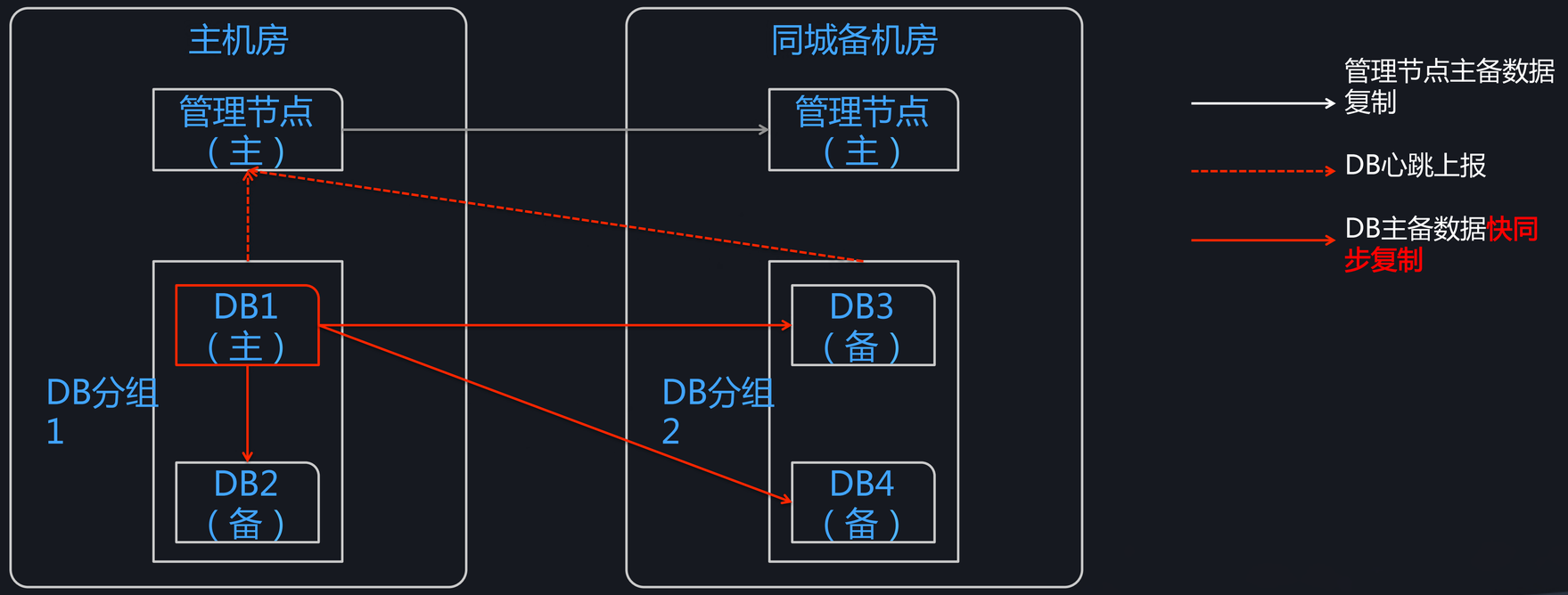


强一致多副本复制技术

实现原理

1、数据高可靠采用DB主备多副本，主备数据复制采用**快同步复制技术**（在MySQL原生半同步复制基础上改造），对DB进行分组管理，要求每个分组至少有一个备机返回响应。

2、服务高可用采用管理节点监控DB状态，管理DB的故障切换。



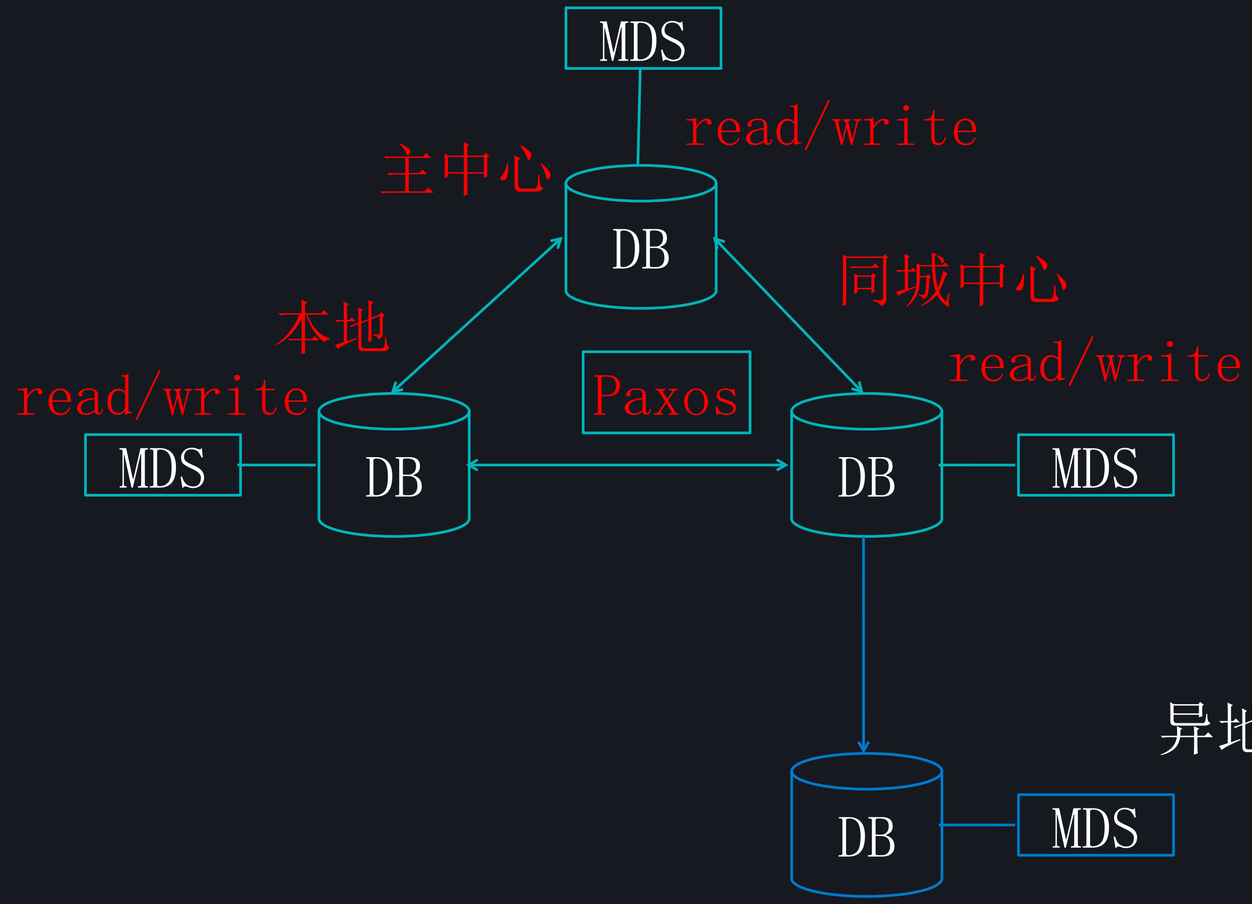
方案特点

1、方案成熟，在同城环境下数据复制性能提升明显；

2、支持DB分组管理，降低单点故障的影响，保障同城机房与主机房数据副本的一致性；

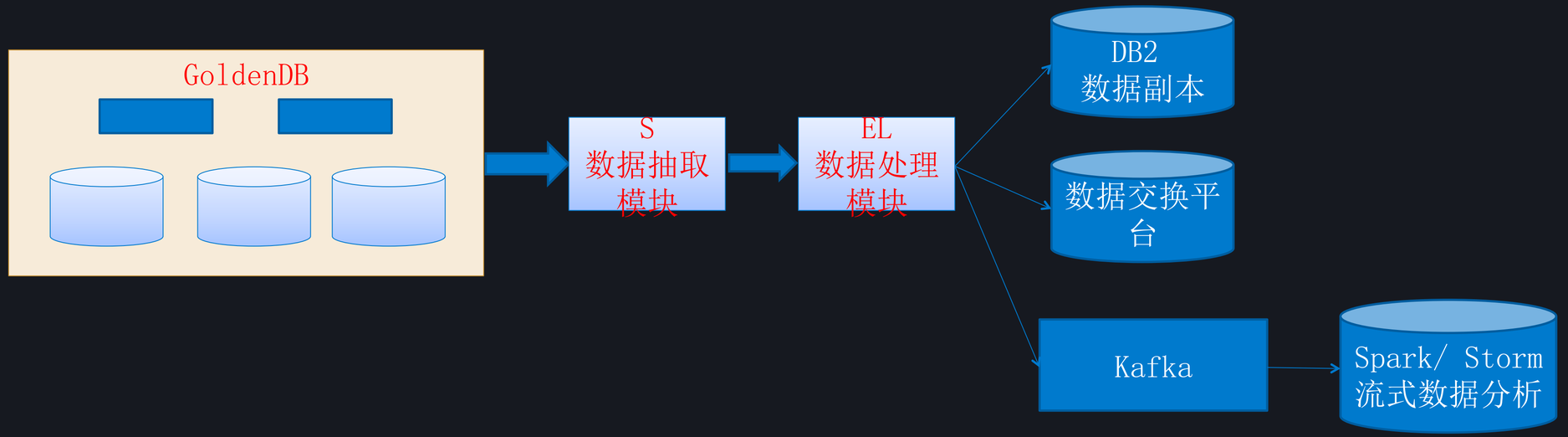
3、支持备机分类，优先在主机房内切换；

元数据一致性



采用Paxos技术,实现元数据存储集群多写强一致

数据迁移一致性



S: Select
数据全量导出
增量数据抽取
条件数据抽取
一致性数据过滤

E:Extract
数据过滤
格式转换
数据合并
数据压缩

T: Transfer
数据传输
数据重传
数据恢复

L: Load
数据转入
DB -> MQ
DB -> DB
DB -> 文件

数据一致性比对工具

提供批量数据比对工具（ 分别批量checksum, 比较checksum值, 不一致则生成不一致数据块单行checksum表进行逐行比对）， 冲突数据根据事先规则进行一致性修复或生成报告交给人工处理。
因目标库表结构与源端异构，需要按映射关系计算CheckSum值

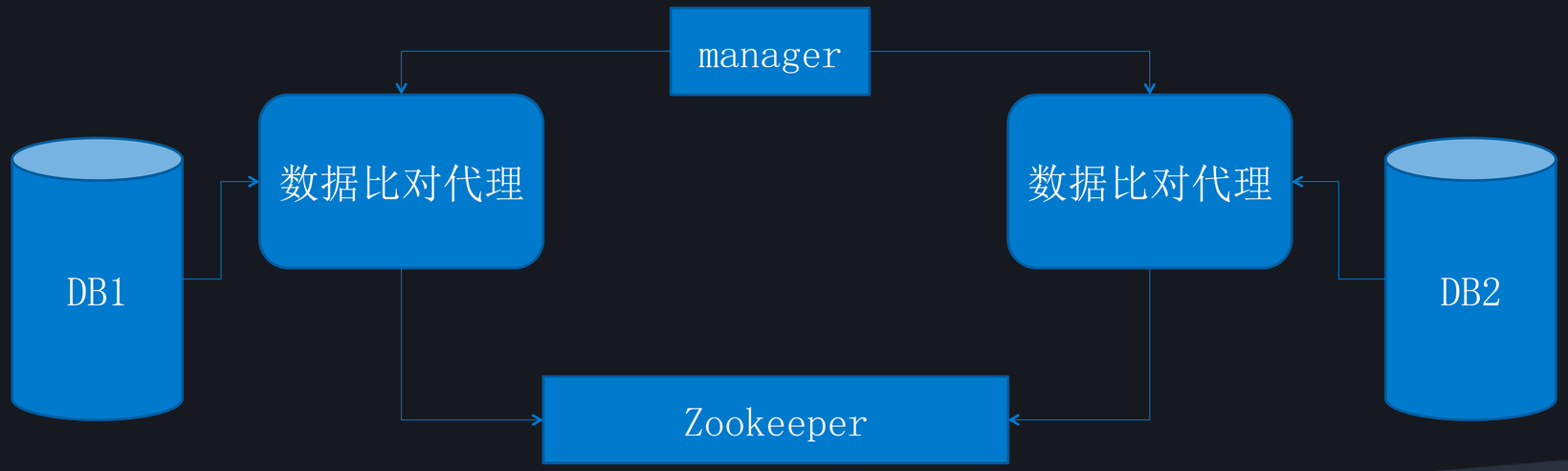
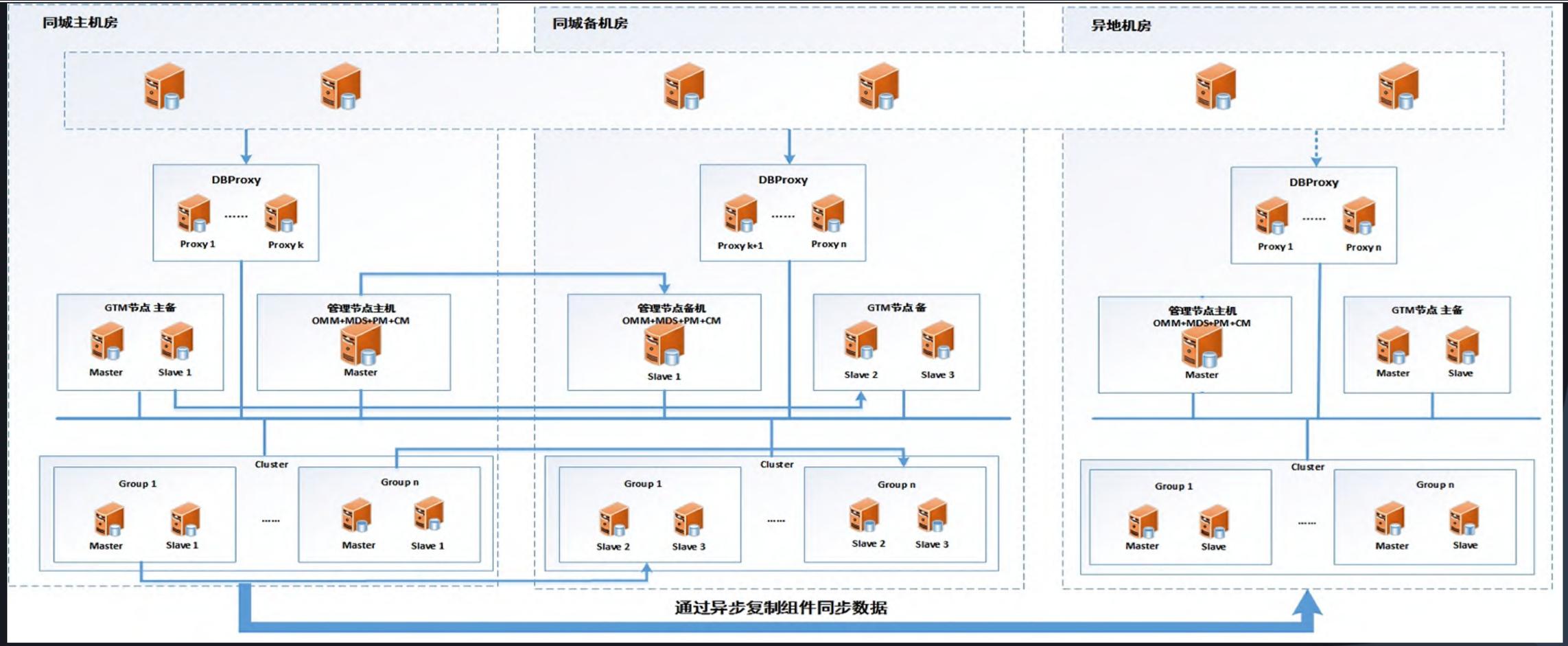


TABLE OF

CONTENTS 大纲

- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：业务连续性
- 挑战3：数据安全性
- 挑战4：服务可扩展性
- 金融行业应用实践

容灾多活架构



RPO=0　RTO<10秒

系统容错设计

- 客户端异常
  - 网络异常
  - 死链接
  - 慢SQL
- Proxy节点异常
  - Proxy当机
  - 与DB断链
  - 存储空间不足
  - 网络故障
- DB节点异常
  - Master 失效
  - 复制失效或降级
  - 存储故障
  - 网络故障
  - 事务状态不一致
- 管理节点异常
  - 元数据失效
  - 非法访问
  - GTM异常
  - PM、CM等当机

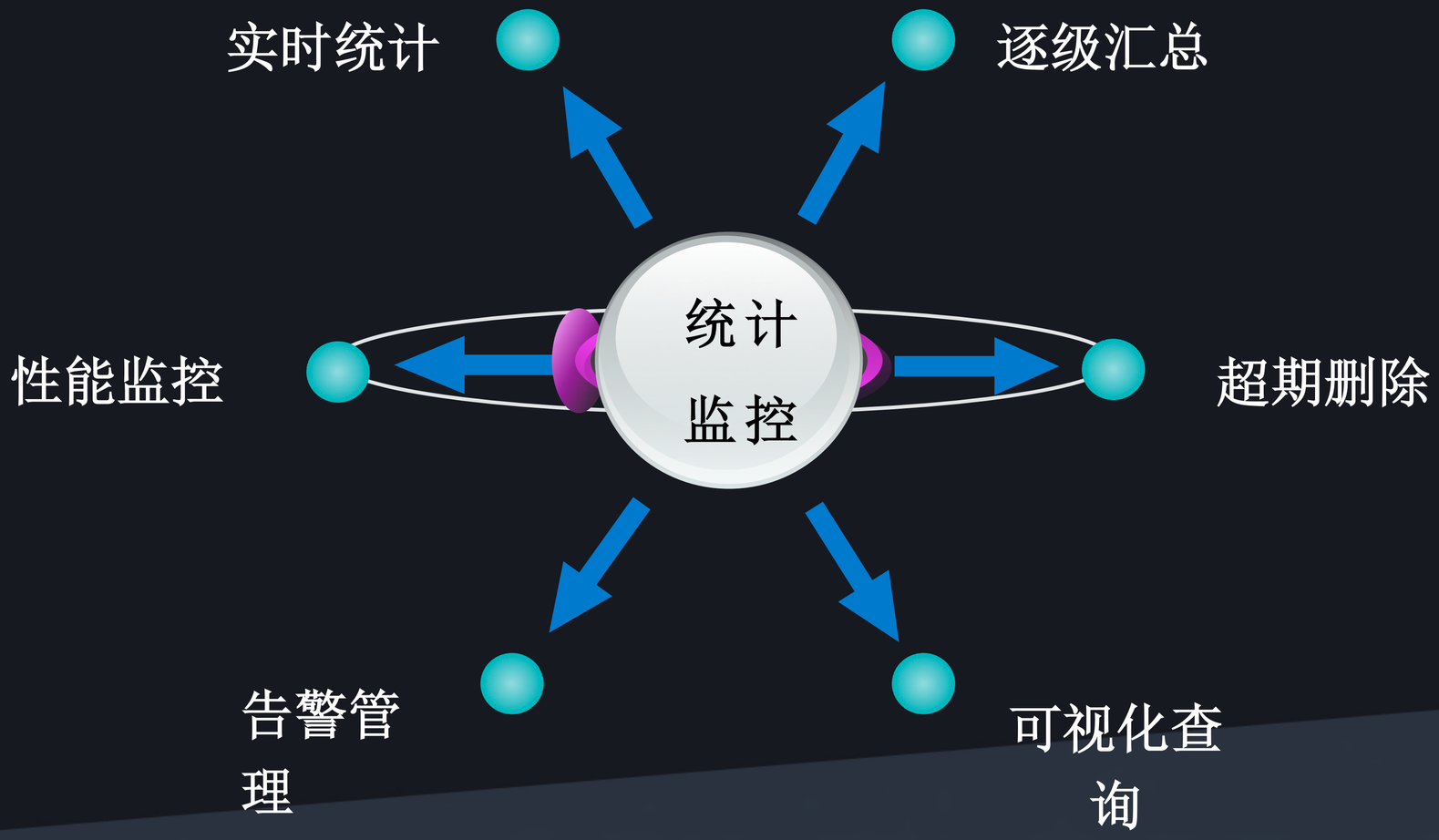
TABLE OF

CONTENTS 大纲

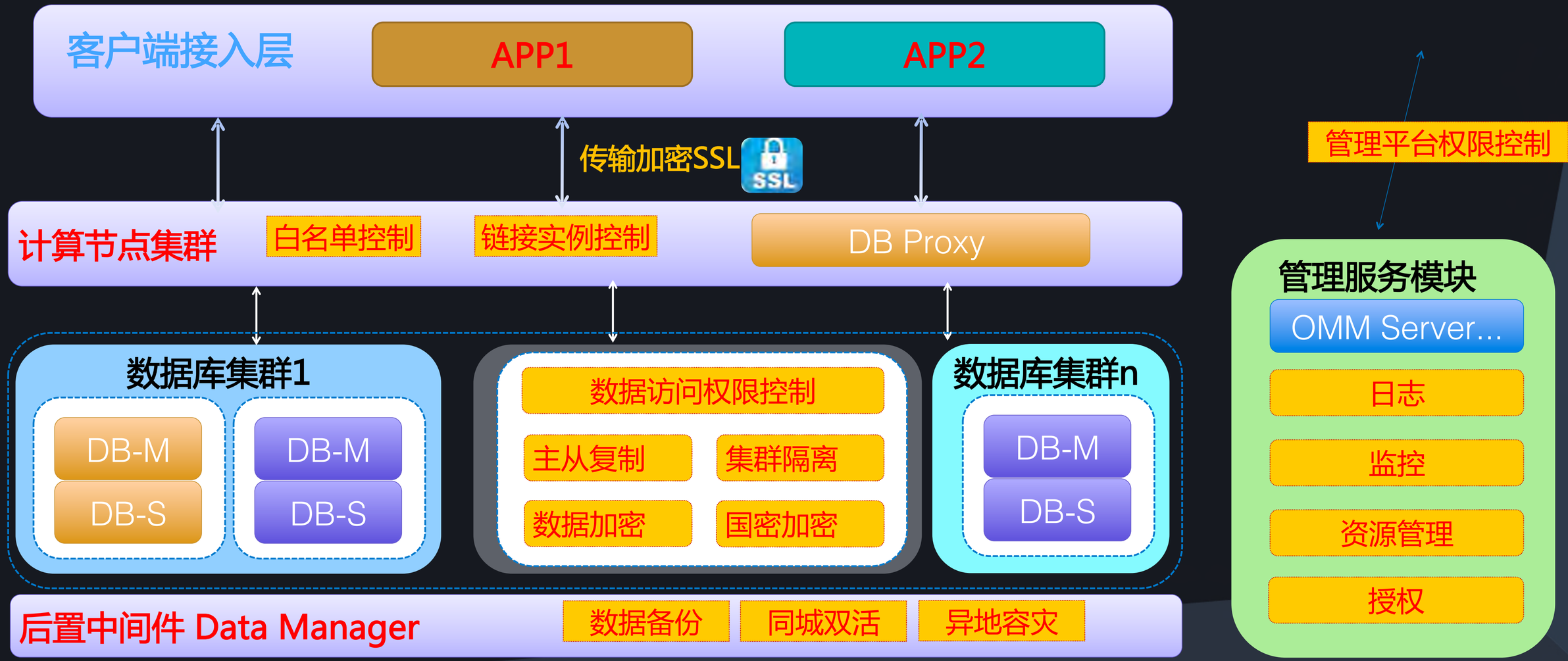
- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：业务连续性
- 挑战3：数据安全性
- 挑战4：服务可扩展性
- 金融行业应用实践

多种数据备份策略

安全运维



数据安全控制


客户端接入层
APP1
APP2
传输加密SSL
SSL
管理平台权限控制
计算节点集群
白名单控制
链接实例控制
DB Proxy
管理服务模块
OMM Server...
日志
监控
资源管理
授权
数据库集群1
DB-M
DB-S
DB-M
DB-S
数据访问权限控制
主从复制
集群隔离
数据加密
国密加密
数据库集群n
DB-M
DB-S
后置中间件 Data Manager
数据备份
同城双活
异地容灾

合规审计

应用欺诈

提取敏感信息

非正常授权

Audit Requirements	COBIT (SOX)	PCI-DSS	ISO 27002	Data Privacy & Protection Laws	NIST SP 800-53 (FISMA)
1. Access to Sensitive Data (Successful/Failed SELECTs)		✓	✓	✓	✓
2. Schema Changes (DDL) (Create/Drop/Alter Tables, etc.)	✓	✓	✓	✓	✓
3. Data Changes (DML) (Insert, Update, Delete)	✓		✓		
4. Security Exceptions (Failed logins, SQL errors, etc.)	✓	✓	✓	✓	✓
5. Accounts, Roles & Permissions (DCL) (GRANT, REVOKE)	✓	✓	✓	✓	✓

网络白名单

数据库异常

SQL防火墙

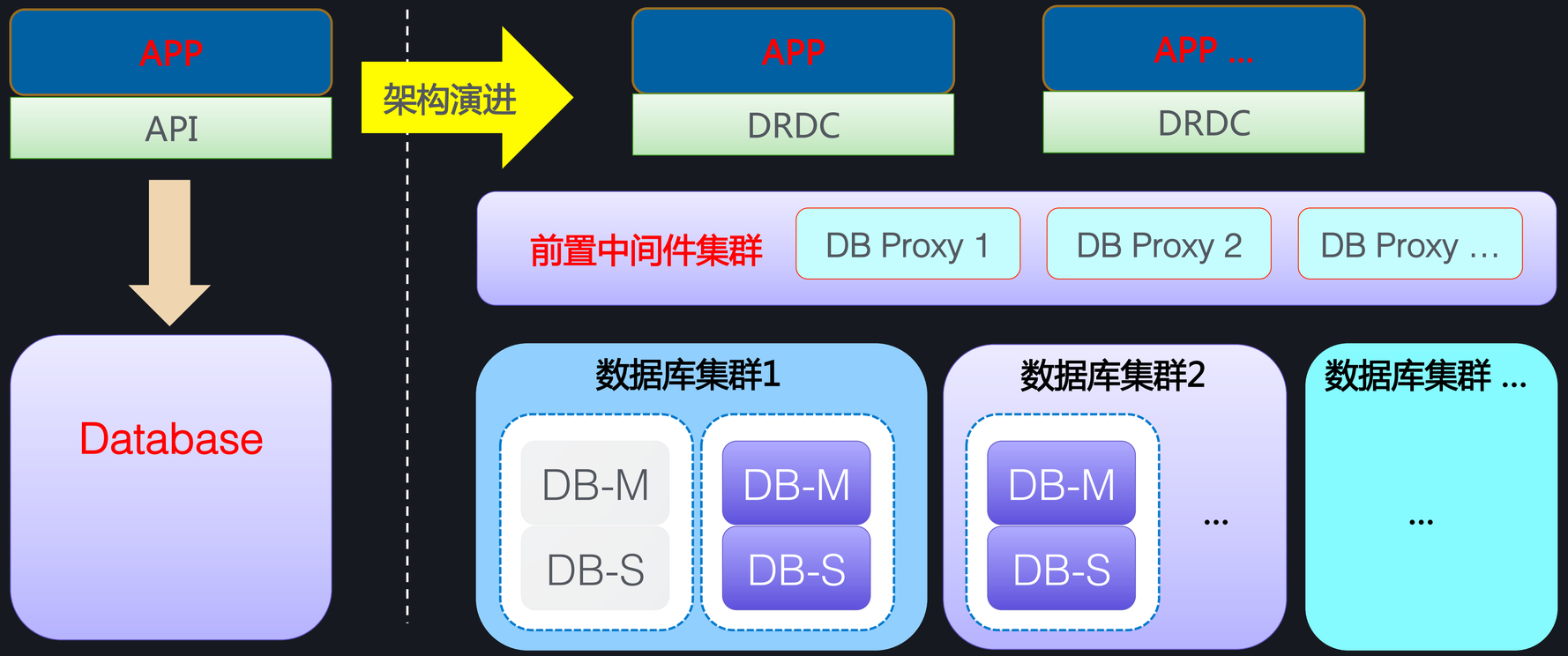
高频查询

TABLE OF

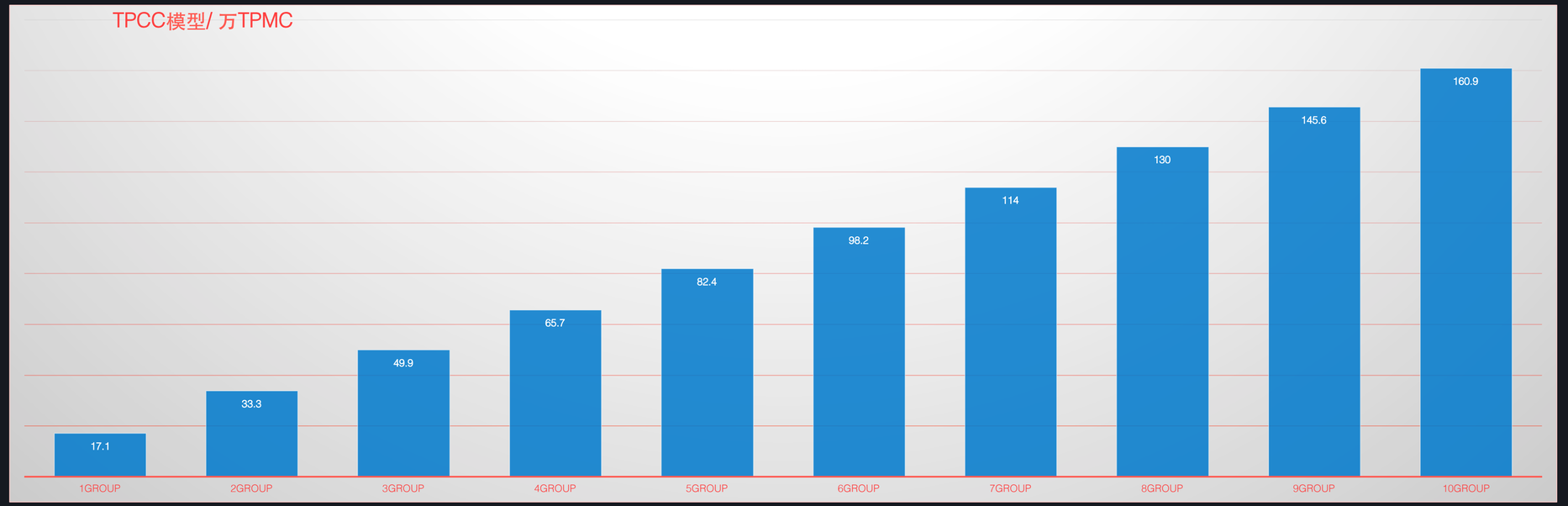
CONTENTS 大纲

- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：业务连续性
- 挑战3：数据安全性
- 挑战4：服务可扩展性
- 金融行业应用实践

容量在线扩容


APP
API
Database
架构演进
APP
DRDC
APP ...
DRDC
前置中间件集群
DB Proxy 1
DB Proxy 2
DB Proxy ...
数据库集群1
DB-M
DB-S
DB-M
DB-S
数据库集群2
DB-M
DB-S
...
数据库集群 ...
...

性能可线性扩展



GoldenDB理论上支持无限大的线性扩展。在TPCC模型下，使用普通X86服务器：

1个安全组的组网下能达到17万；

5个安全组的组网下能达到82万；

10个安全组的组网下能达到160万；

性能近似线性，衰减少于7%。

93%

正比例扩展能达到93%以上的性能

多种读写隔离级别提升性能

读语句级别

UR(uncommitted read)：未提交读，即不判断分布式读写冲突，适用于允许脏读或者不存在读写冲突的业务场景；

CR(consistency read)：强一致性读，先查询活跃GTID，后查询数据，严格保证返回结果处于分布式事务已提交状态，不存在脏读的可能性；

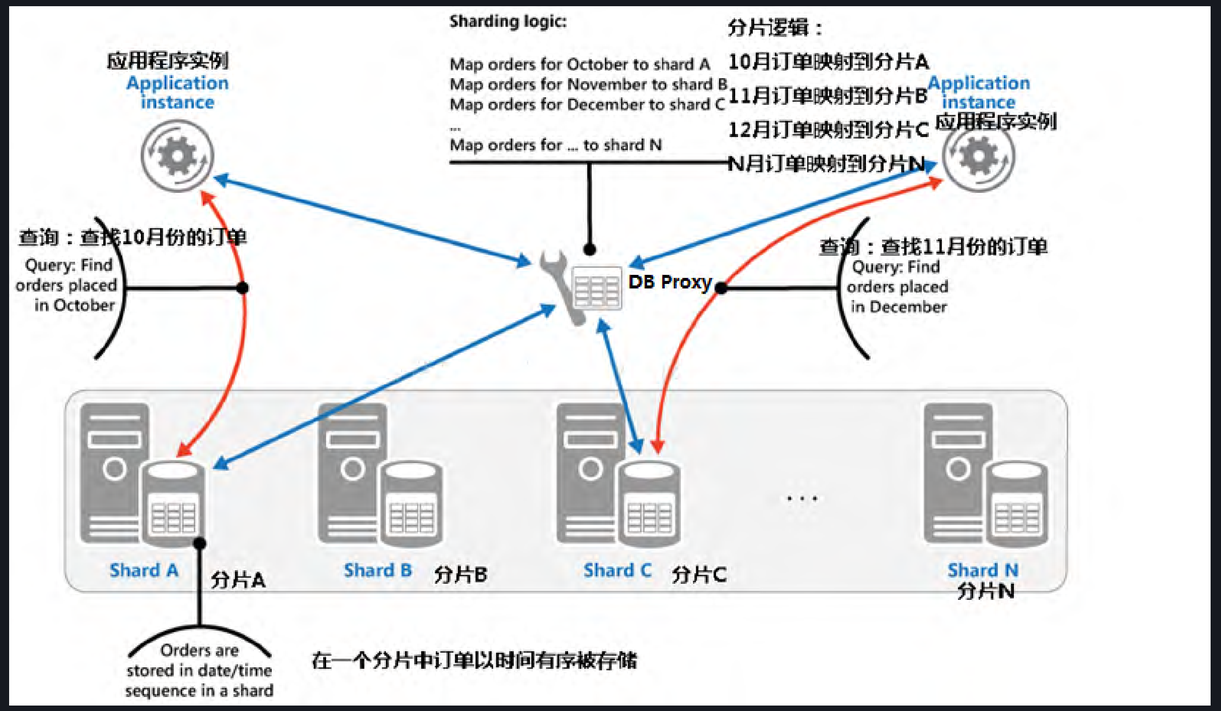
SEMI-CR(semi- consistency read)：半强一致性读，同时查询活跃GTID和数据，仅判断GTM中的活跃事务，在高并发读写时存在极小概率的脏读，但效率较CR高；

写语句级别

SW(single write)：单事务写，即不判断分布式写写冲突，适用于不存在多个事务同时写相同数据的场景；

CW(consistency write)：强一致性写，需要判断分布式写写冲突，允许多个事务同时写相同的数据；

灵活数据分片




```
Create table bank.info (Customer_Number int key,
Corporate_Property varchar(30), Private_type int,
Corporate_Information varchar(30) distributed by
case Corporate_Property
when '中信银行伦敦' then g9
when '中信银行纽约' then g10
else
        case Private_type
        when '对私' then subdistributed by hash(Customer_Number)(g1,
g2, g3, g4, g5)
        else
                case Corporate_Information
                when '五矿集团' or '光大集团' then g6
                else
                        subdistributed by hash(Customer_Number)(g7, g8)
                end
        end
end
```

支持按特定规则多重分片，满足复杂业务需求

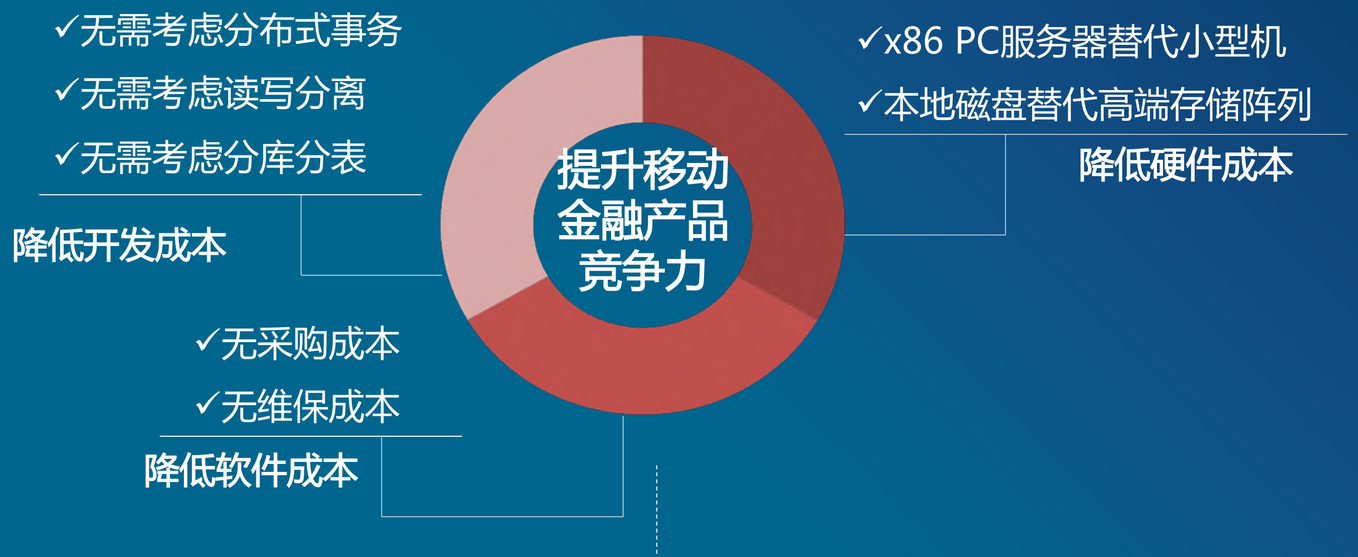
TABLE OF

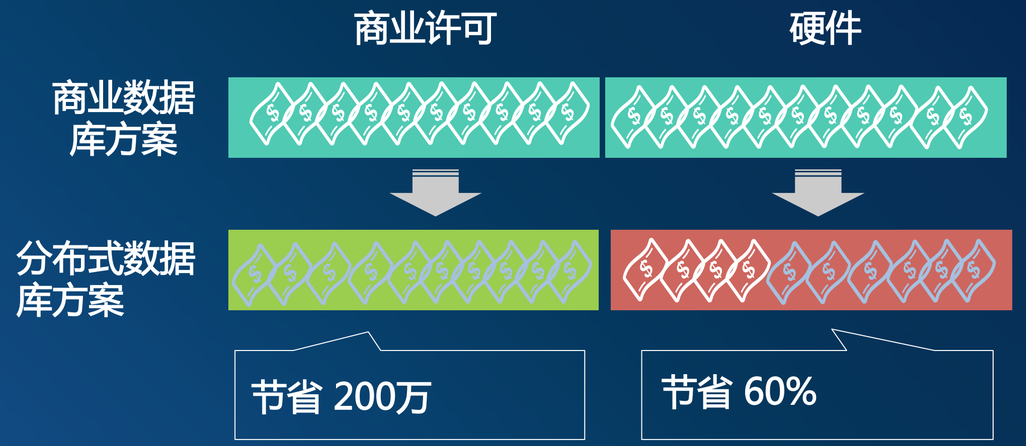
CONTENTS 大纲

- 金融数字化趋势与挑战
- 挑战1：分布式数据一致性
- 挑战2：业务连续性
- 挑战3：数据安全性
- 挑战4：服务可扩展性
- 实践案例

中信银行


2014年
启动分布式数据库研发
分布式数据库1.0问世
2015年7月
2015年9月
冠字号系统合作平台上线
中信银行门户网站系统上线
2016年5月
2016年11月
金融同业平台成功上线
总行核心下移验证
2017年4月
2017年6月
零售客户综合积分系统上线


✓无需考虑分布式事务
✓无需考虑读写分离
✓无需考虑分库分表
降低开发成本
✓x86 PC服务器替代小型机
✓本地磁盘替代高端存储阵列
降低硬件成本
提升移动金融产品竞争力
✓无采购成本
✓无维保成本
降低软件成本


商业许可
硬件
商业数据库方案
分布式数据库方案
节省 200万
节省 60%

ZTE中兴

江苏银行事后监督系统案例

江苏银行事后监督系统采用GoldenDB替换现在使用的Oracle数据库，提升系统处理性能。

项目情况简介

1.2017年4月上线。

2.性能情况：典型存储过成执行速度提升3倍（Oracle VS 3节点GoldenDB集群）。

3.2 Proxy /3 Group/ 10 虚拟机（2P16C/64G/ SATA/万兆网卡）

系统特点介绍

1.业务介绍：由影像处理、业务监督、辅助功能三部分组成。

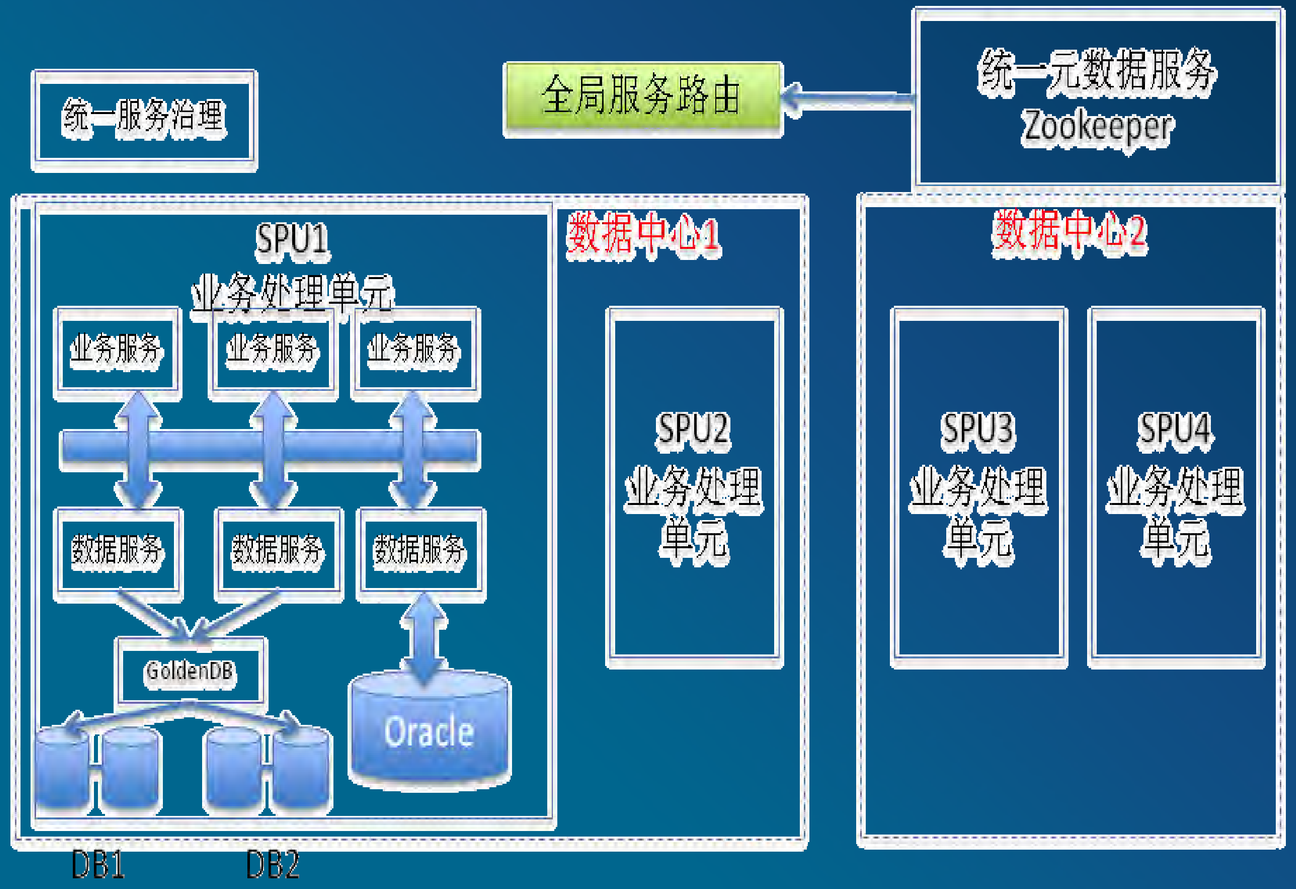
2.改造工作：现有系统使用Oracle数据，已运行7年，涉及403张表及200多个存储过程。

3.性能容量要求：现有数据量2T，目标规划10T的数据容量。采用3节点安全组，每晚导入9G的数据进行跑批处理。

4.关键功能：GoldenDB的存储过程能力。

某银行项目综述

1.GoldenDB目前在某行开展的工作包括综合积分业务对接和行方分布式平台应用改造二个项目。
2.已经完成基准测试、TPCC性能测试、手机银行业务对接测试。



综合积分业务对接

1.背景：行方期望通过综合积分系统进行分布式架构试水，成功后推广到其他业务。
2.业务介绍：共6个业务子模块，原系统部署在16个Oracle RAC集群上， 目标是实现GoldenDB 与Oracle同时混合部署。
3.容量：共7000万用户，GoldenDB对接测试使用3个DBGroup集群
4.关注功能点：Oracle兼容性、数据安全性及数据一致性

分布式平台应用改造

1.项目介绍：综合项目，涉及分布式数据库、大数据及Pass平台等。
2.里程碑：17年5月底完成分布式平台开发与验证工作。

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全 球 架 构 师 峰 会 2017