华为Cloud Native分布式数据库技术

Calvin Sun, Huawei Cloud BU

May 11, 2018

Calvin Sun

MySQL®

2006.01

ORACLE®

INNODB®

2008.02

HUAWEI

2017.10

ORACLE®

Cloud

2016.02

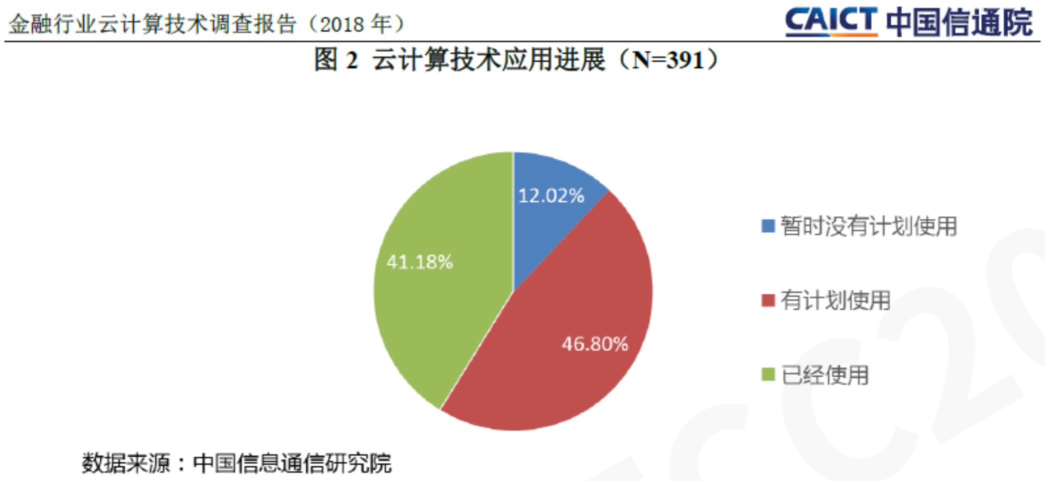
2013.03

Agenda

- Cloud Computing in Fintech
- Cloud Native Database
- Huawei Cloud Native Distributed Database System

Cloud Computing in Fintech

金融行业云计算技术



金融行业云计算技术调查报告（2018 年） CAICT 中国信通院

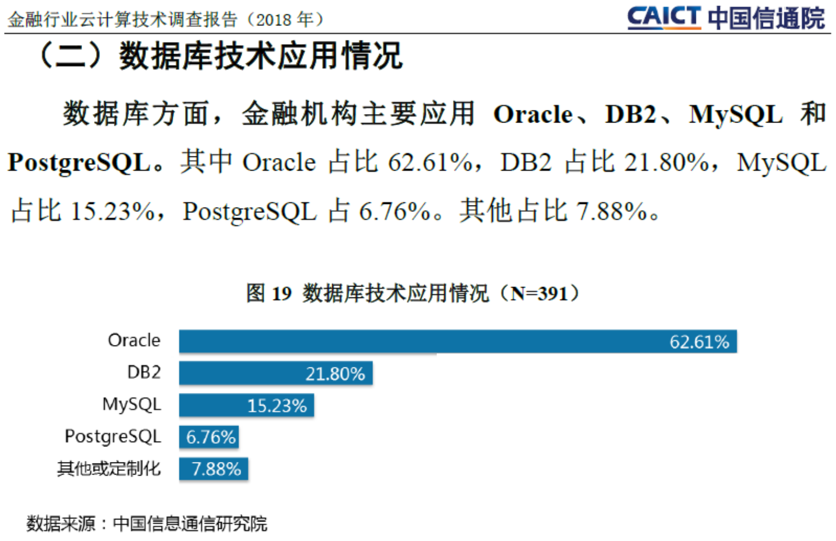
图 2 云计算技术应用进展（N=391）

数据来源：中国信息通信研究院

百花齐放春满园，ArkDB俏枝头
——MySQL金融级解决方案的百家争鸣时代

周彦伟 极数云舟 CEO, 05/11 15:30 – 16:10

演讲简介：随着MySQL的广泛应用和深入人心，对数据库技术要求最高的金融领域也慢慢开始转型到MySQL，随之而来的就是业内各种各样的MySQL金融级解决方案百花齐放，百家争鸣。本次分享首先是对业内的一些做法做了调研，同时也详细讲述了完全基于开源引擎Galera Cluster的MySQL金融级解决方案的实现方式ArkDB，它最大的特点是:基于开源引擎Galera、可自主控制、高一致性、高安全性、久经考验，在完全满足金融系统要求的同时，也大大增加了系统自身的可控性，是金融系统的又一个重要选择。



金融行业云计算技术调查报告（2018 年） CAICT 中国信通院

（二）数据库技术应用情况

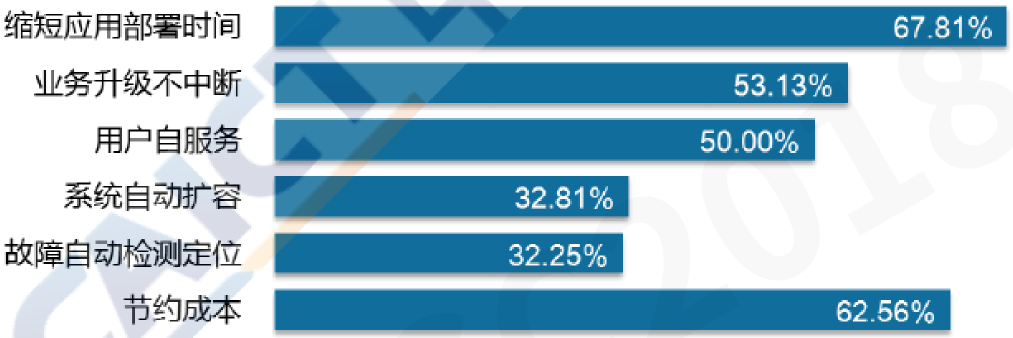
数据库方面，金融机构主要应用 **Oracle**、**DB2**、**MySQL** 和 **PostgreSQL**。其中 Oracle 占比 62.61%，DB2 占比 21.80%，MySQL 占比 15.23%，PostgreSQL 占 6.76%。其他占比 7.88%。

图 19 数据库技术应用情况（N=391）

数据来源：中国信息通信研究院

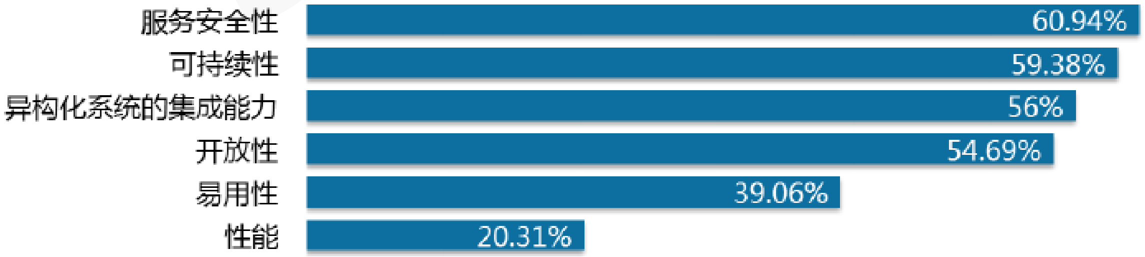
金融行业云计算技术

图 3 用户应用云计算技术的目的（N=344）



金融行业云计算技术调查报告（2018 年）　　CAICT 中国信通院

图 4 用户对云计算技术的要求（N=344）

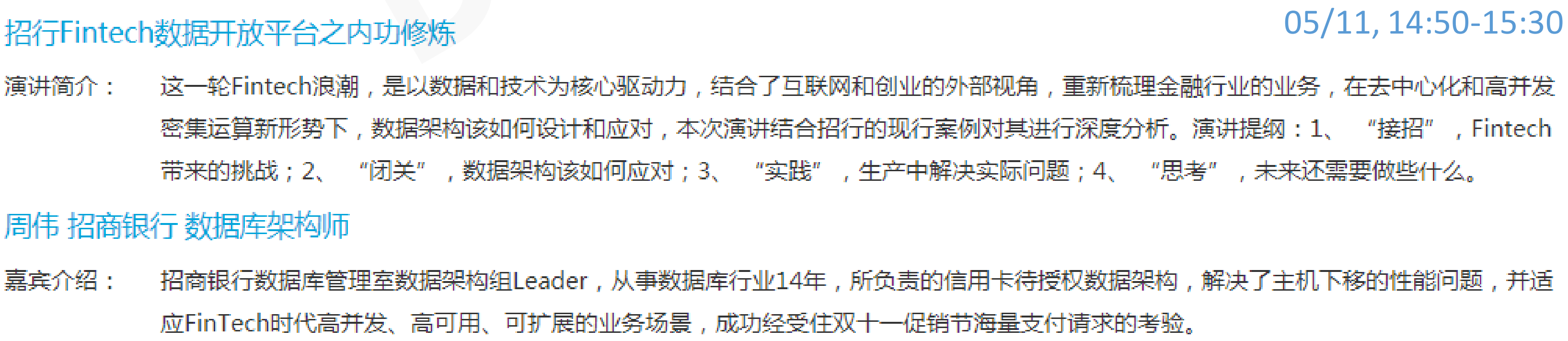


数据来源：中国信息通信研究院

华为与金融行业云技术

招商银行和华为成立分布式数据库联合创新实验室，实践云上科技变革

【中国，深圳，2017年12月4日】2017年11月27日，华为与招商银行股份有限公司在招商银行研发中心举行了分布式数据库联合创新实验室揭牌仪式。双方将共同应对“CloudFirst”的挑战，利用云、大数据、人工智能先进技术，领先的金融业务实践和优秀资源，联接业务与技术，联合进行分布式数据库技术的研发和产品应用，解决数据库应用上云的问题。华为云助力招商银行加速数字化转型，成为“金融科技银行”，通过科技变革，为客户提供普惠、个性化、智能化的金融服务。



招行Fintech数据开放平台之内功修炼

05/11, 14:50-15:30

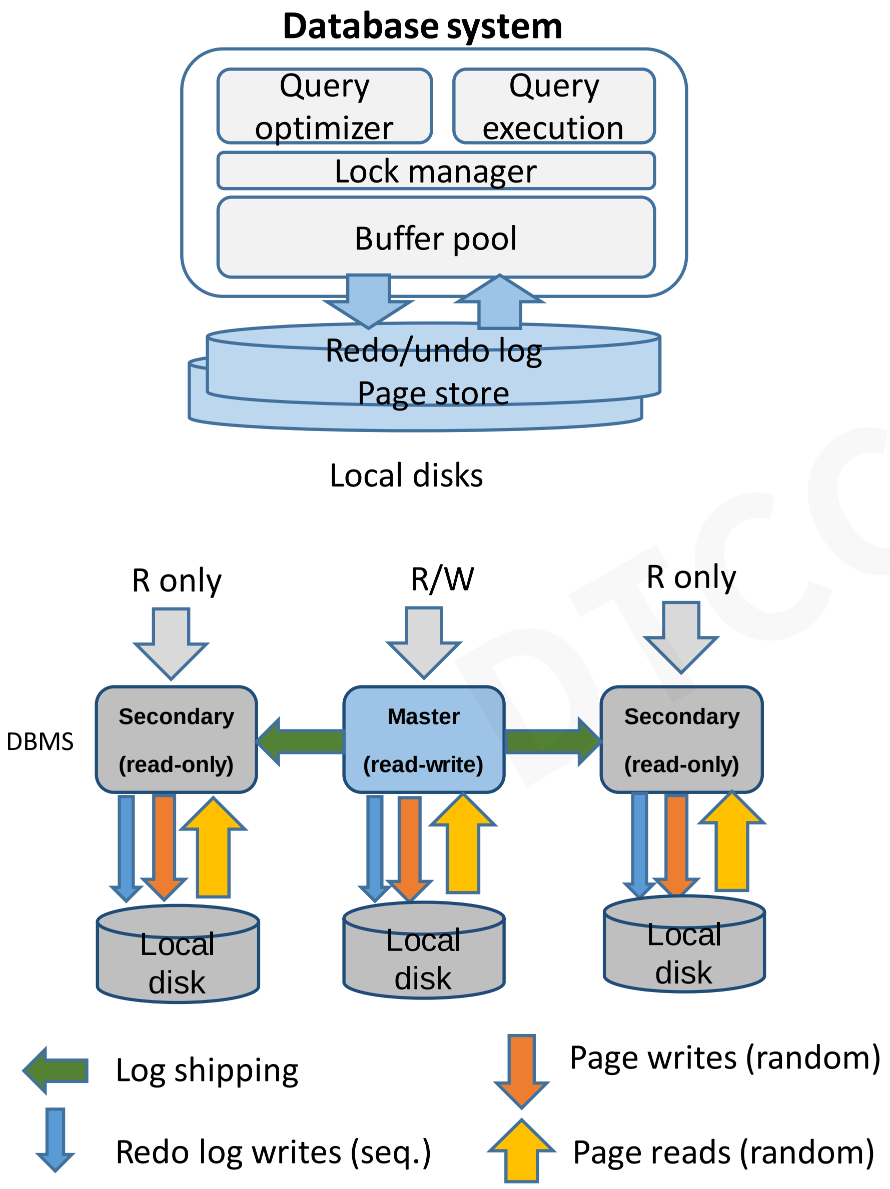
演讲简介： 这一轮Fintech浪潮，是以数据和技术为核心驱动力，结合了互联网和创业的外部视角，重新梳理金融行业的业务，在去中心化和高并发密集运算新形势下，数据架构该如何设计和应对，本次演讲结合招行的现行案例对其进行深度分析。演讲提纲：1、 “接招” ，Fintech带来的挑战；2、 “闭关” ，数据架构该如何应对；3、 “实践” ，生产中解决实际问题；4、 “思考” ，未来还需要做些什么。

周伟 招商银行 数据库架构师

嘉宾介绍： 招商银行数据库管理室数据架构组Leader，从事数据库行业14年，所负责的信用卡待授权数据架构，解决了主机下移的性能问题，并适应FinTech时代高并发、高可用、可扩展的业务场景，成功经受住双十一促销节海量支付请求的考验。

Cloud Native Database

Traditional RDBMS with HA



- Basic architecture is 30+ years old
- Designed for hardware at the time (single processor, small memory, slow hard disks)
- Master plus two secondaries for reliability (HA)
- Master handles all updates and ships log to each secondary
- Each secondary updates its own copy of the database
- If master fails, a secondary becomes the new master
- Secondaries may handle read-only queries
- Log shipping is only network load

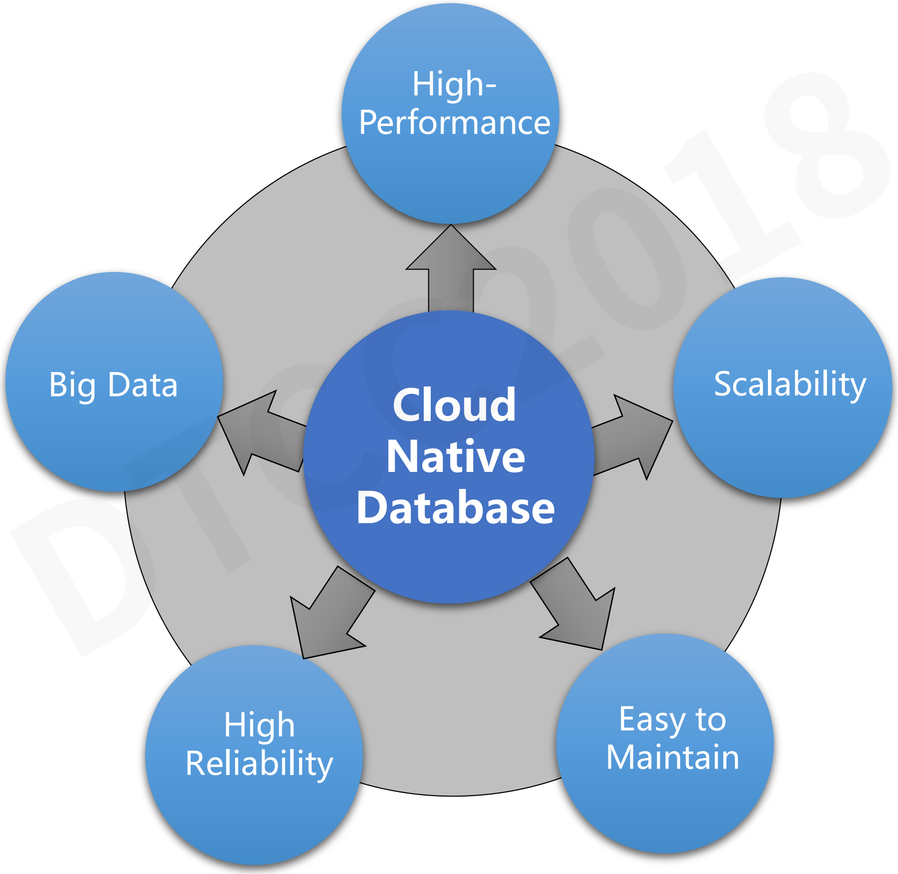
Trends in Cloud Databases

- Separation of compute and storage

  Gartner: *By 2019, 90% of cloud DBMS architectures will support separation of compute and storage, rendering those that do not as irrelevant in the overall market.*

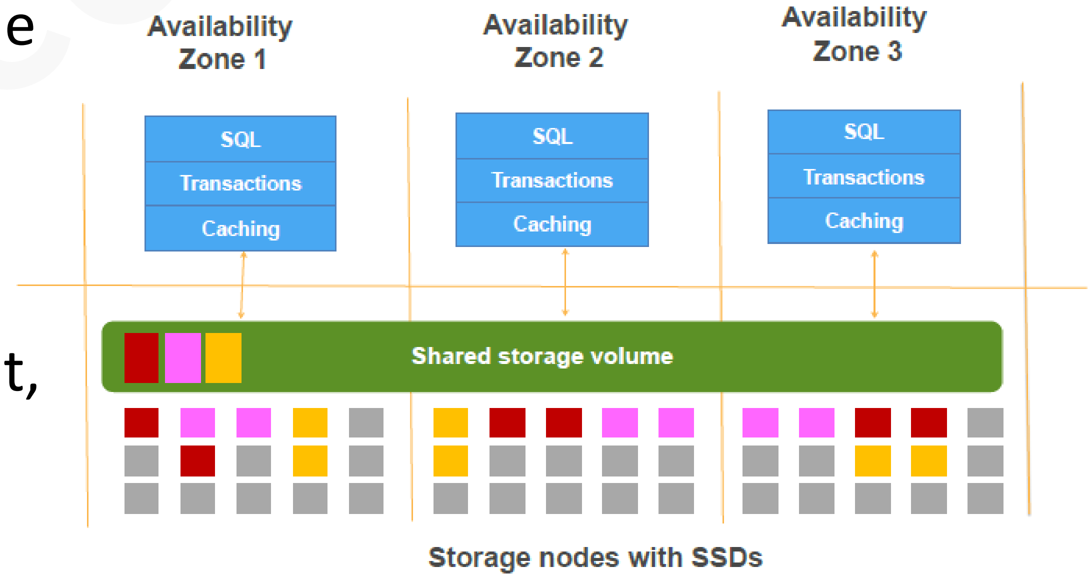
- How to leverage latest hardware advances:
  - CPU: Multi-cores with NUMA
  - Storage: Optane SSDs (Coldstream & AEP)
  - Network: RDMA
  - Special hardware: GPU, FPGA
- Advances in AI and ML

Requirements of Cloud Native Database



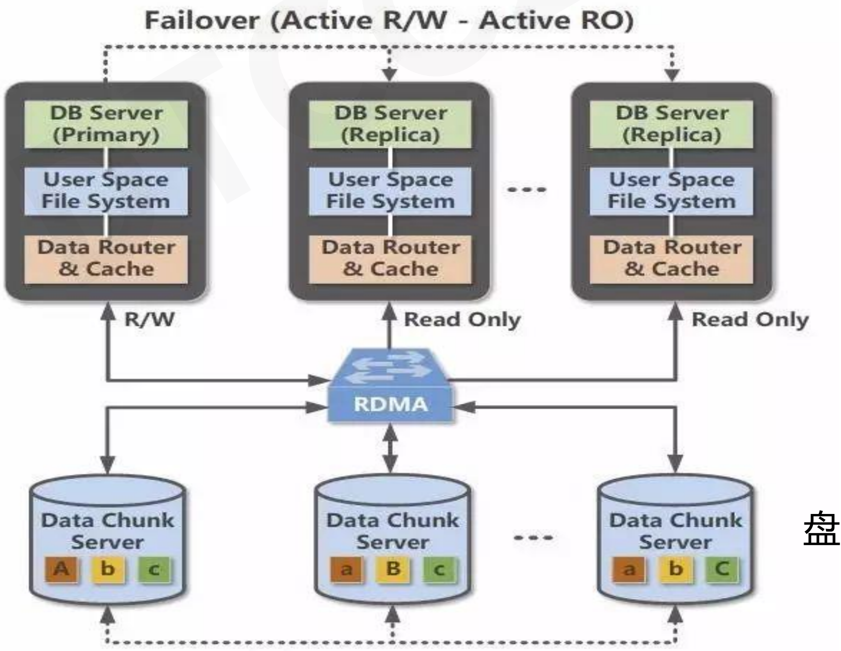
AWS Aurora Highlights

- Separation of storage: allows compute and storage to be scaled separately
  - Compute layer is responsible for maintaining ACI(no 'D').
  - Storage layer ensures Durability.
    - Storage service receives redo logs; coalesce; and periodically apply the redo logs to data pages
    - Near-instantaneous crash recovery
    - Replicate data locally (within the storage node)
- Leverage redo log across the distributed storage system
  - Significantly reduces network traffics.
  - Efficient append-only logs.
- Improved lock management, query cache, etc.



PolarDB Highlights

- Separation of compute and storage
- Redo-based physical replication
- RDMA for fast redo transfer
- RDMA based optimization for parallel-raft
- User-mode file system: optimized for redo and page IO
- 3D XP + NVMe
- Many SQL node optimizations (MySQL 5.6 based)



盘古

Huawei Cloud Native Database System

Huawei Cloud Native Database Design Principles (1)

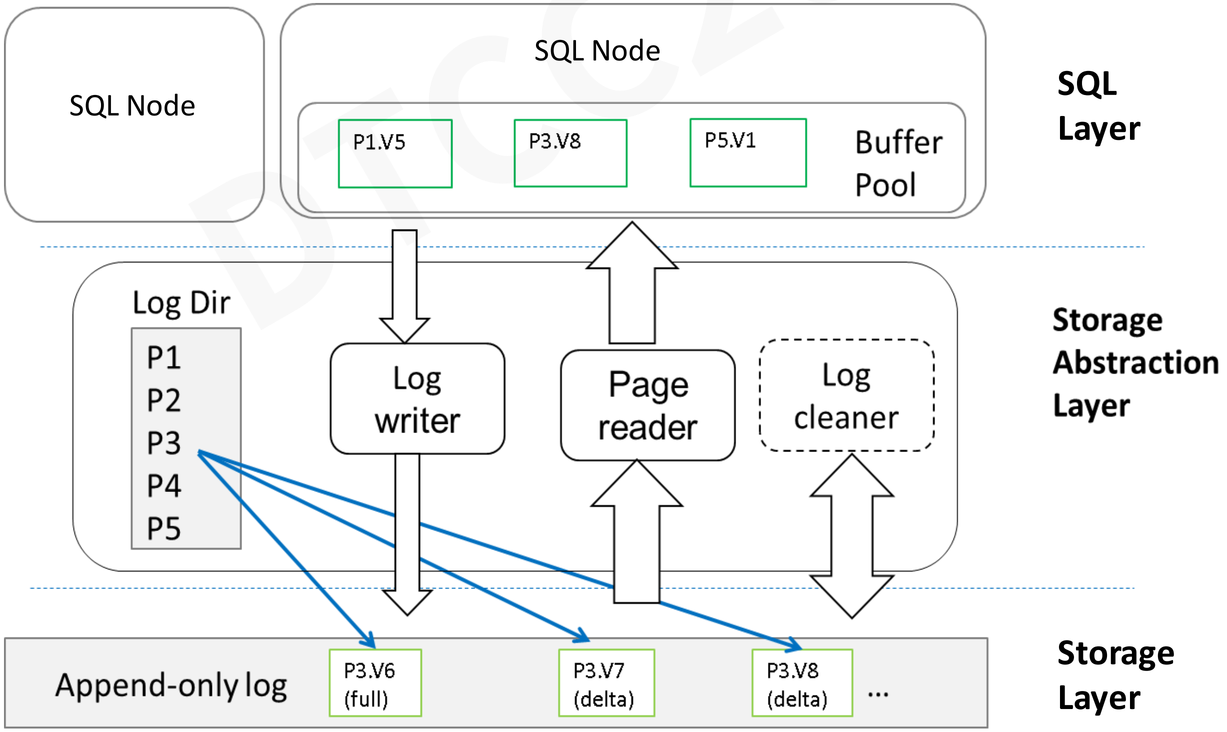
- Decoupling:
  - Separate compute from storage
  - Decouple master from replica
- Pushdown operations close to data
  - Offloading work to storage nodes, such as redo processing, page construction
- Exploit functionality provided by cloud storage
  - Building storage as an independent fault-tolerant and self-healing service
  - Shared access (single writer, multiple readers)
- Exploit properties of SSDs
  - Avoid random writes to SSDs to minimize wear
  - Exploit good random read performance of SSDs

Huawei Cloud Native Database Design Principles (2)

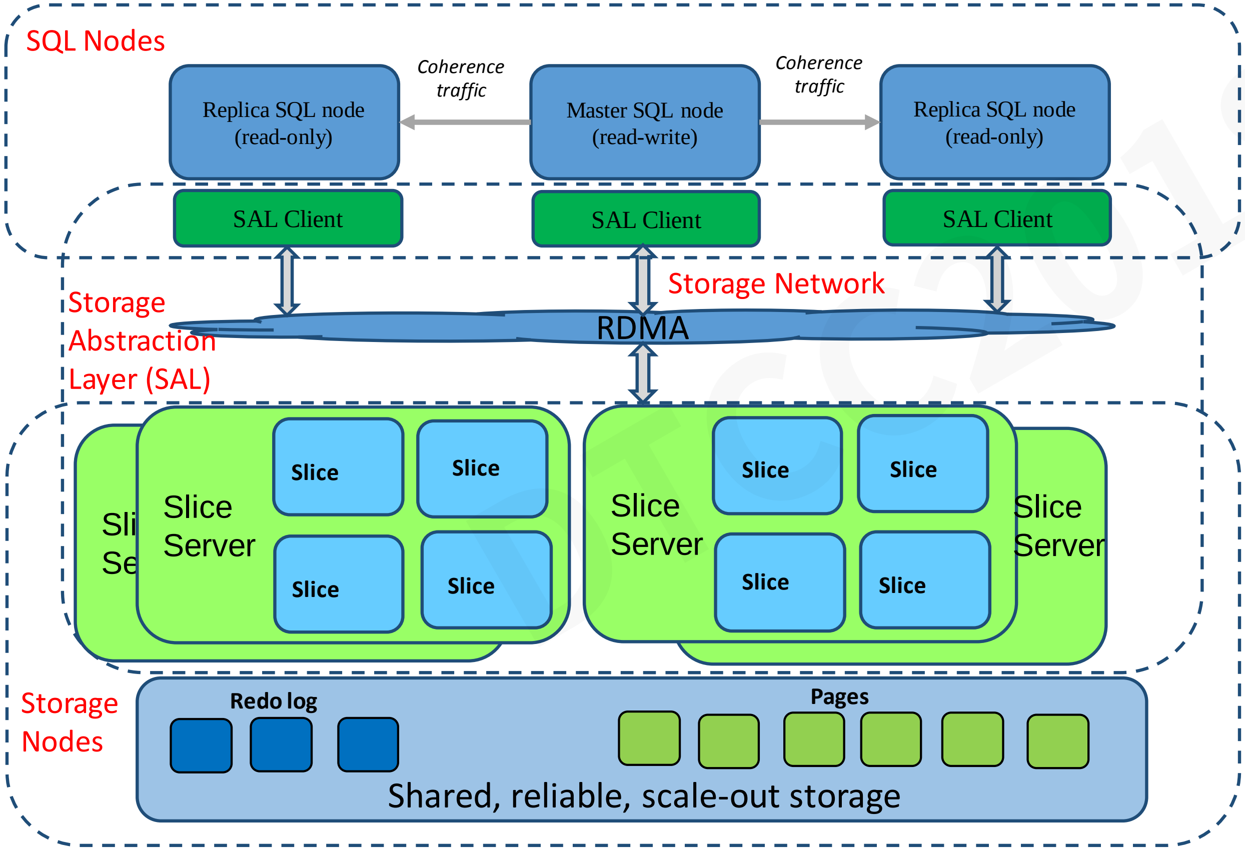
- Multi-tenant support
- Believe the central constraint in high throughput data processing has moved from compute and storage to the network.
  - Reduce network traffics
  - Take advantages of new network technologies, e.g. RDMA
- Leverage advances in AI and ML for autonomous system
  - Auto-scaling, self-tuning

High-level Architecture

- Separation of compute and storage
  - Easy scale-out and load balancing
- Storage abstraction layer (SAL) isolates SQL front-end, transaction and query execution from the way storage is organized
  - Foundation for additional database support
- Storage layer provides fast, reliable, shared storage
  - Use storage in a log structured manner; only sequential writes to minimize SSD wear



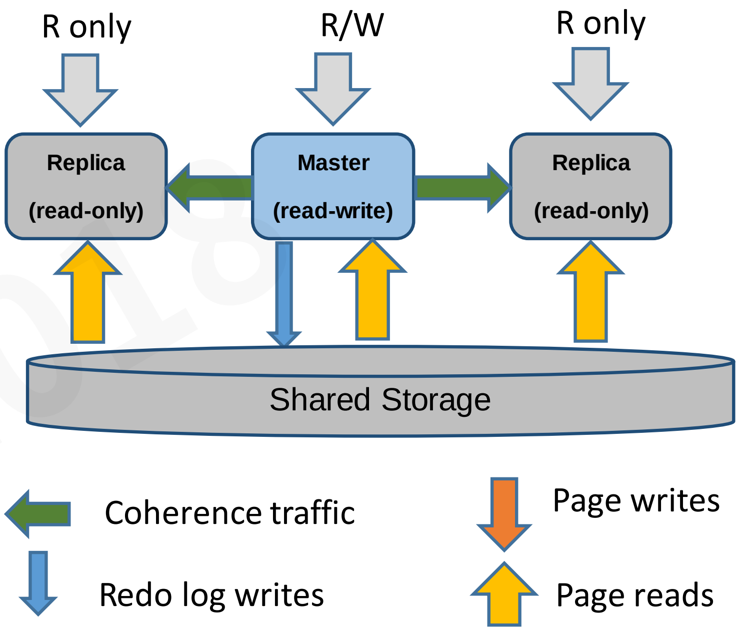
Overview of Huawei Cloud Native Database



- Master database server
  - Handles all updates
  - Writes to the WAL logs
- Read Replica database servers
  - Can handle read-only requests
  - Enable fast failover
  - Can be added at any time
- Database data is partitioned across storage nodes
  - Pages are logically organized based on slice and distributed among slice servers
  - Each slice is duplicated for reliability
  - Log records for a page are sent to the corresponding slice
- Slice Server
  - Maintain multiple slices for different tenant databases
  - Store and process log records
  - Maintain and construct pages
  - Serve page read requests

SQL Nodes

- Managing client connections, parsing SQL requests, planning and executing queries and managing transaction isolation
- One RW with multiple RO replica
- Loose coupling between master and replica
- Light traffic between master and replica
- Enable fast failover
- Build on top of HWSQL with additional enhancements
  - Query result cache
  - Query plan cache
  - Online DDL



MySQL云数据库的性能优化和发展趋势

05/11, 15:10-16:00

演讲简介： MySQL是当前最流行的开源数据库，高性能是各个云数据库提供商的核心竞争力。在这次讲座中，将简要回顾MySQL关键性能特征，并介绍MySQL云服务的现状，以及我们为提高华为云上的MySQL的性能所做的工作。最后我们还将讨论进一步改进MySQL在云上性能的一些挑战和机遇。

SQL Replica

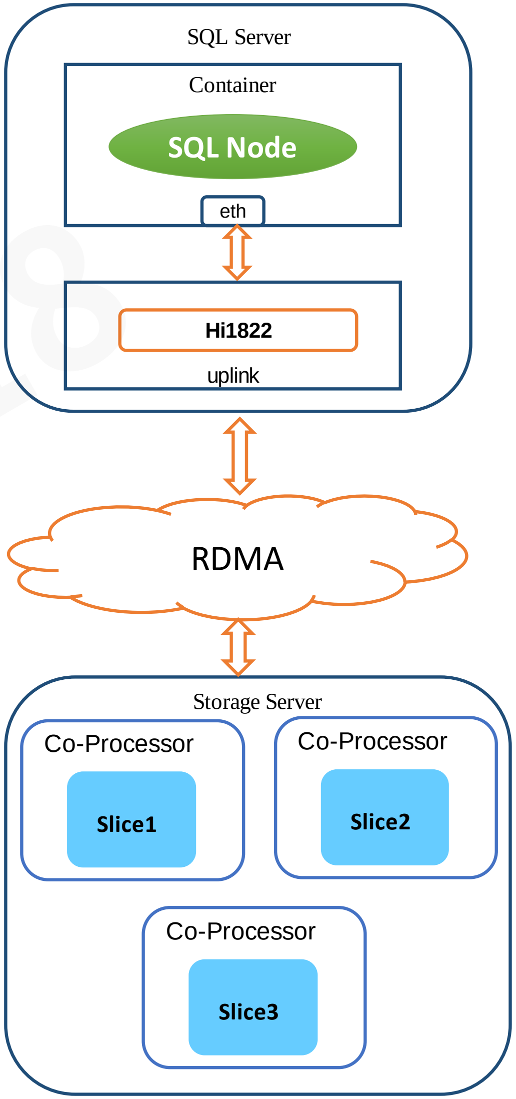
- Replica maintains multiple versions of page in its buffer pool
  - Reduce frequent page read from storage.
- Interactions with Master
  - For MVCC: read replicas will receive the list of currently active transactions, ("readview“), from the master periodically.
  - For page invalidation: as updates happen on the master, read replicas also receive the list of pages updated during transaction commits.
  - For purge: read replicas need to feed their min LSNs to the master. The master conducts the purge operation based on info from all read replicas.
  - For DDL operations: master coordinates with read replicas (via MDL) when performing DDL operations.

Storage Abstraction Layer (SAL)

- Storage abstraction layer (SAL) is a logical layer
- Isolates SQL front-end, transaction and query execution from the way storage is organized
- Consists of a common log module executing on SQL nodes and slice stores executing on storage nodes.
- SAL operates on database pages and supports access to multiple versions of the same page.
- SAL divides all data pages into slices, based on {spaceID, pageID}.
  - Scaling out – as the database grows in size, resources (storage, memory) available to grow proportionally as more slices are created.
  - Data locality – data intensive operations executed by slice servers running on the storage nodes that store the data.

Performance Enhancements

- Container with Hi1882 Chip
  - Lower network latency
  - Higher network PPS
  - Less cost of system software
- RDMA
  - Lower latency between SQL and storage
  - Support thousands nodes
- Co-Processor
  - Near Data Process
  - Help to reduce SQL node workload



Summary

- Separation of compute and storage with a logical storage abstraction layer (SAL)
- Exploit functionality provided by cloud storage
  - HA features: atomic write, replication, failover, …
  - Shared access (single writer, multiple readers)
- Exploit properties of SSDs
  - Avoid random writes to SSDs to minimize wear
  - Exploit good random read performance of SSDs
- Multi-tenant support
- Take advantages of new network technologies, e.g. RDMA
- Pushdown operations close to data
  - Offloading work to storage nodes
- Leverage advances in AI and ML for autonomous system

THANKS

HANA
Hadoop
SQL
MySQL
DB2
BigData
Spark
SQL Server
MPP
NewSQL
Oracle
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168