HUAWEI

MySQL云数据库的性能优化和发展趋势

Calvin Sun, Huawei Cloud BU

May 11, 2018

Calvin Sun

HUAWEI

MySQL®

2006.01

ORACLE®

INNODB®

2008.02

HUAWEI

2017.10

ORACLE®

Cloud

2016.02

2013.03

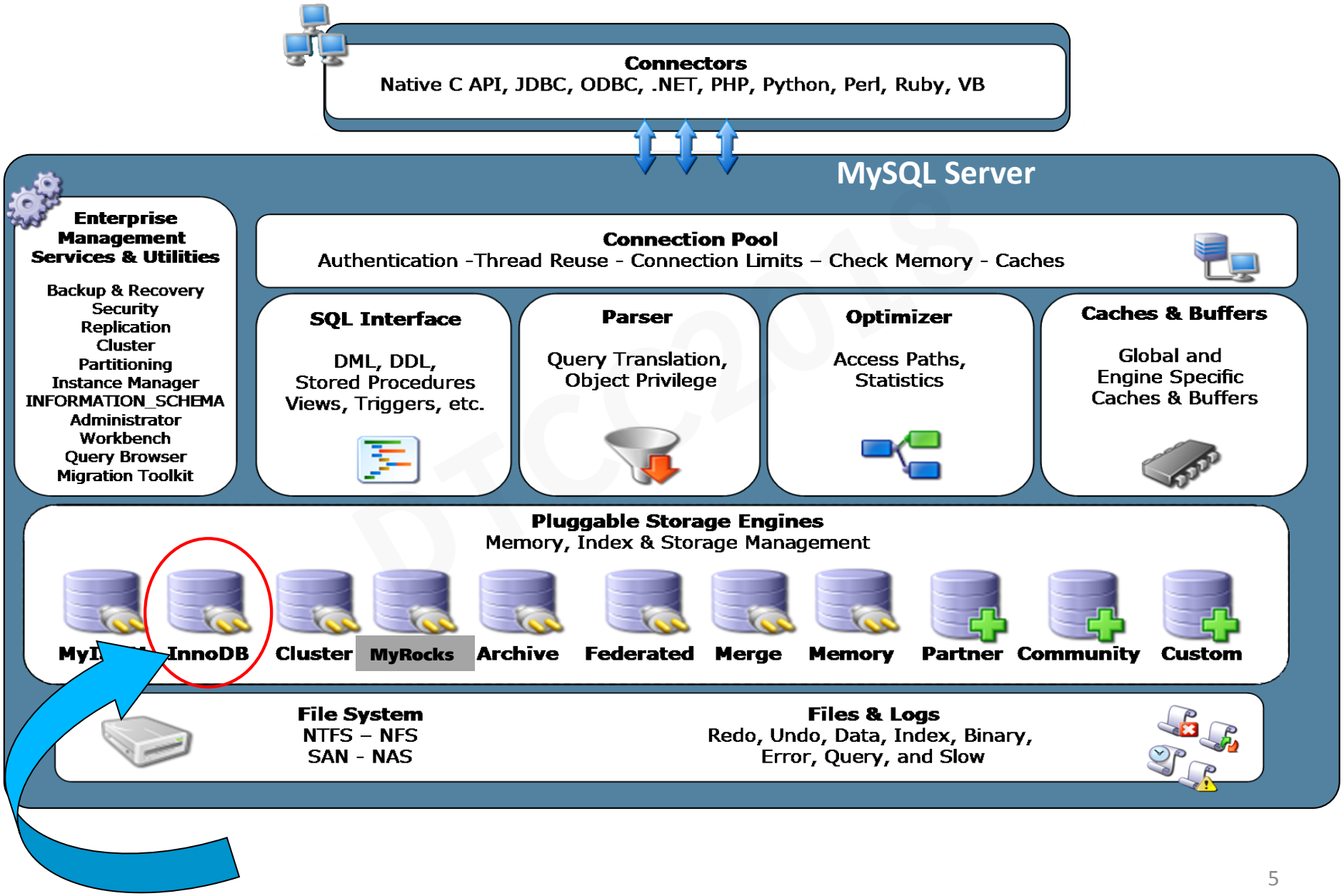
Agenda

- MySQL Key Performance Features
- Huawei RDS MySQL Family & Performance Improvements
- Challenges & Opportunities

HUAWEI

MySQL Key Performance Features

MySQL Server Architecture



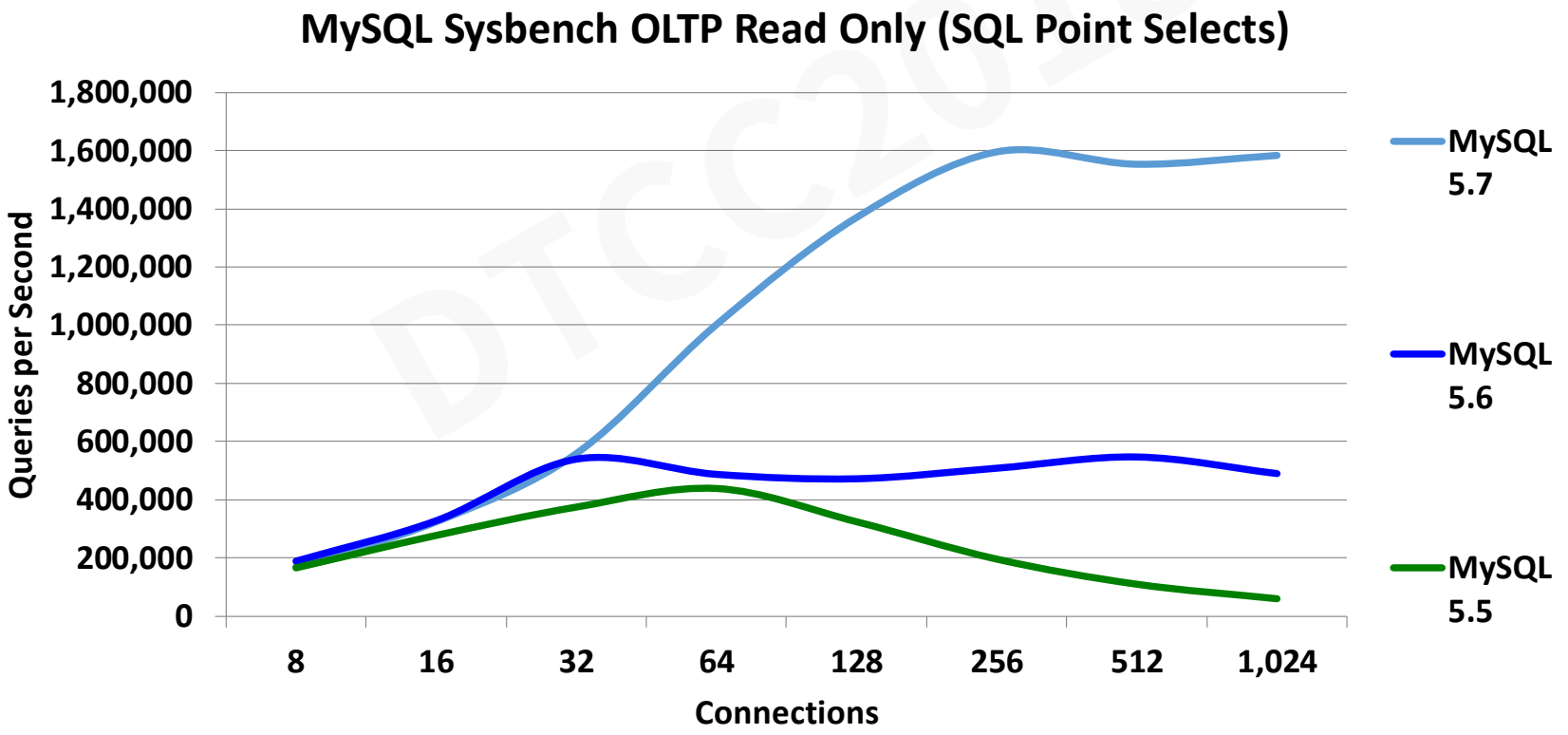
MySQL / InnoDB Performance Features

- Multi-Threaded Architecture
- MVCC with 2PL
- Group Commit
- Adaptive Flushing
- Purging
- Pre-Fetching
- Adaptive Hash Indexes
- Change Buffering

MySQL Sysbench : SQL Point Selects/sec

MySQL 5.7: 3x Faster than MySQL 5.6
MySQL 5.7: 4x Faster than MySQL 5.5

1,600,000 QPS



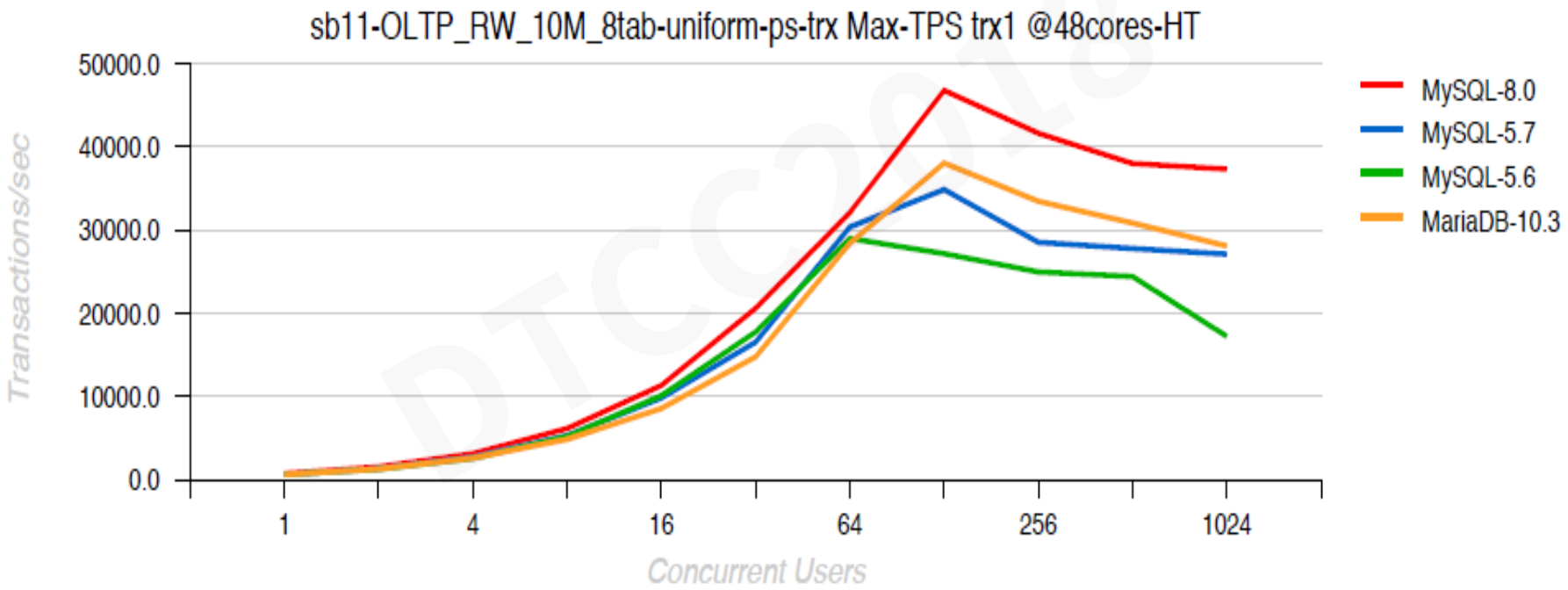
Intel(R) Xeon(R) CPU E7-8890 v3 4 sockets x 18 cores-HT (144 CPU threads) 2.5 Ghz, 512GB RAM Linux kernel 3.16

OLTP_RW latin1 @MySQL 8.0 GA

HUAWEI

MySQL 8.0: 30% Faster than MySQL 5.7

MySQL 8.0: 50% Faster than MySQL 5.6



45K (!!) TPS Sysbench OLTP_RW 10Mx8tab, trx_commit=1, 2S

HUAWEI

Huawei RDS MySQL Family & Performance Improvements

RDS Cloud Services

Out-of-the-Box
Obtain a production-ready relational database in minutes with a few clicks

Easy to Scale
Scale up CPU, memory and storage or deploy multiple read replicas for better transaction throughput

Security Protection
VPCs, subnets, security groups, SSL protections and audit logs

Data Migration
Easily bring data from external sources into Cloud RDS

Backup & Restore
Automated backups, point-in-time restores, snapshots

High Availability
Highly available database services within or across availability zones

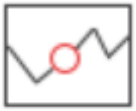
MySQL Family on Huawei Cloud

Product	Description
RDS MySQL	A value-added, cloud-based implementation of the MySQL Server community edition
RDS HWSQL 5.6 (Q1 2018)	An enhanced MySQL Server 5.6 for superior performance and availability, engineered by Huawei Cloud BU
RDS HWSQL 5.7 (Q2 2018)	An enhanced MySQL Server 5.7 for superior performance and availability, engineered by Huawei Cloud BU

Huawei RDS HWSQL

Engineered by Huawei for superior performance and availability

Delivers all the capabilities of Huawei RDS *plus*:



High-Performance

Approximately 3x the performance of MySQL Community Edition

Enhanced Reliability

Enhanced, loss-less semi-sync and semi-sync notification improves reliability and avoids possibility of data loss on standby takeover

Improved Scalability

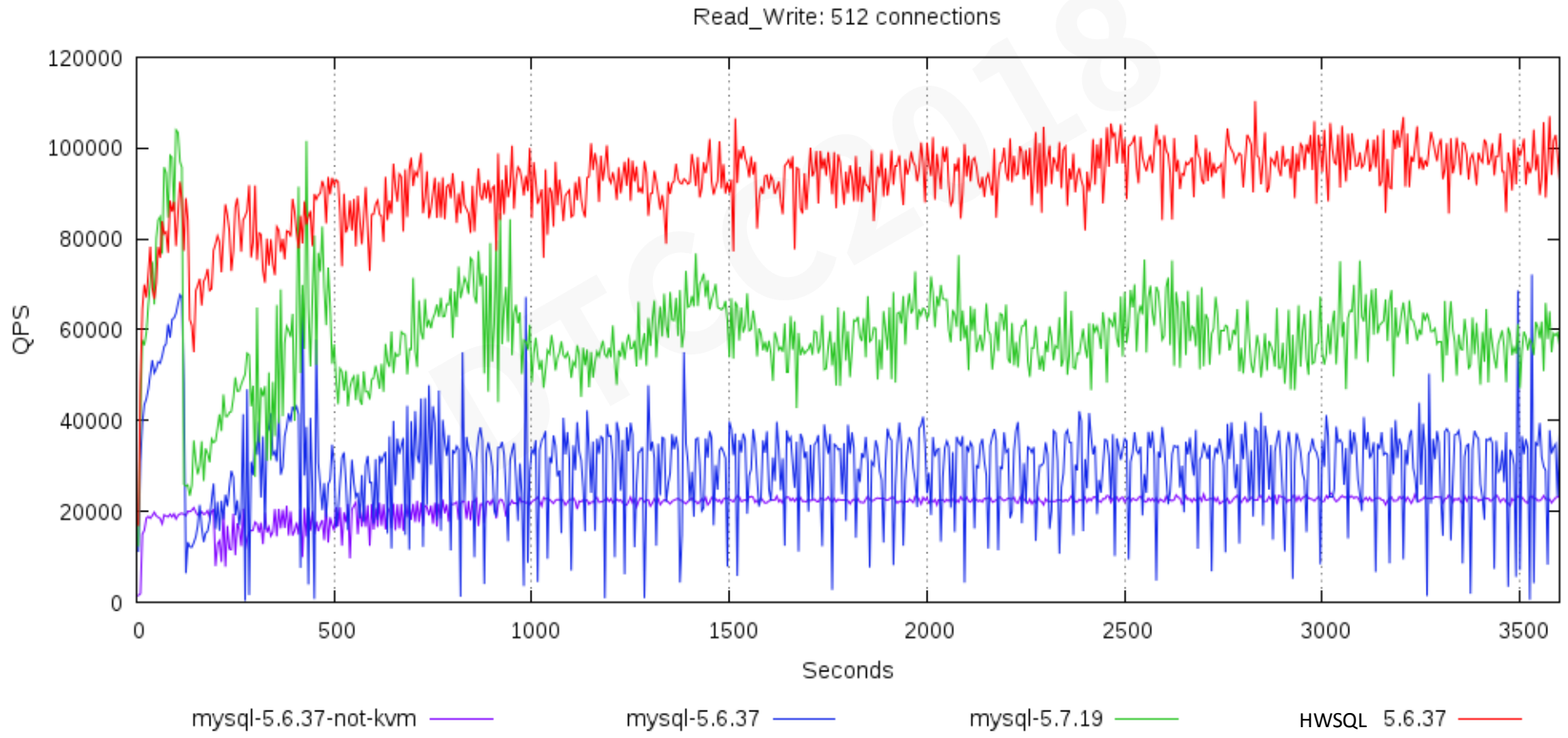
Supports more database clients, more concurrent transactions, and larger server configurations

Fast Recovery on Failure

Improved replication greatly reduces time to recover on failure and data lag on read replicas

Huawei RDS HWSQL: Throughput

- Up to 3x the throughput of community version MySQL server
- Superior, consistent performance with large numbers of clients



HUAWEI

MySQL 5.6 : Performance & Scalability Issues

- High-concurrency workloads performance degradation
  - > 1000 clients → big performance drop
  - Especially on modern large multi-core systems
- Master-Slave replication lag/delay
  - Severe lag for single-database replication
  - MTS replication in 5.6 only works on *different* databases
- Multi-core scalability bottlenecks
  - Workloads cannot scalable on multi-core systems (e.g., 32 cores) in read-heavy, write-heavy, and read-write mixed workloads

HWSQL 5.6: Key Performance Features (1)

- Improve high-concurrency workloads performance
  - Thread-pool plugin
- Solve master-slave replication lag
  - Transaction-level multi-threaded slave (MTS) parallel replication
- Implement loss-less semi-sync replication
  - Make semi-sync master faster
  - Implement the loss-less semi-sync to prevent potential data loss

HWSQL 5.6: Key Performance Features (2)

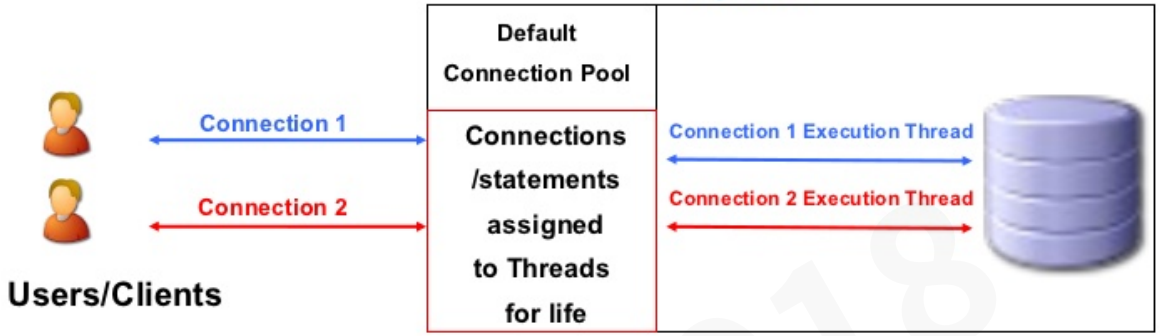
- Multi-core scalability
  - Read-only transactions optimizations
    - Streamlining read-only transactions processing
    - MVCC readview reuse
    - Adaptive Hash Index (AHI) latch enhancement
  - Write transactions & logging optimizations
    - Group commit stages & notification
    - Redo log writing & flushing
    - Buffer pool LRU list scanning
  - Scalable memory manager integration & deep tuning

HWSQL 5.7: Key Performance Features

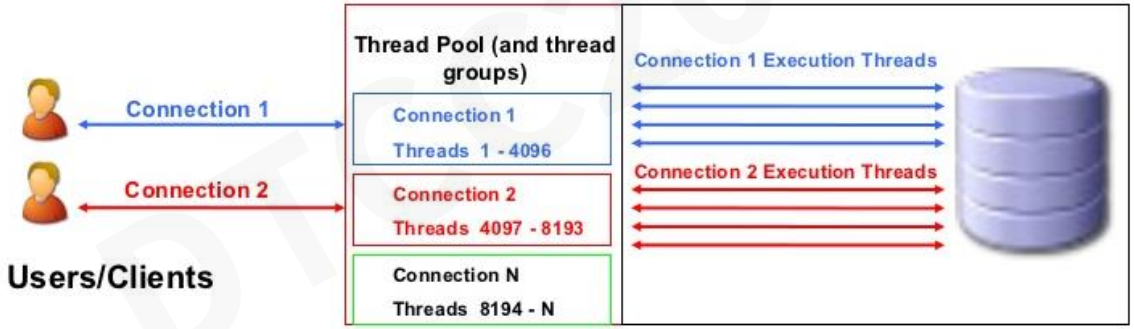
- All in HWSQL 5.6
- SQL aggregate pushdown
  - Pushdown query aggregate evaluation to storage engine to reduce overheads
- Query cache
  - Scalable query cache with auto cache deactivation

Thread-Pool Plugin

MySQL Community



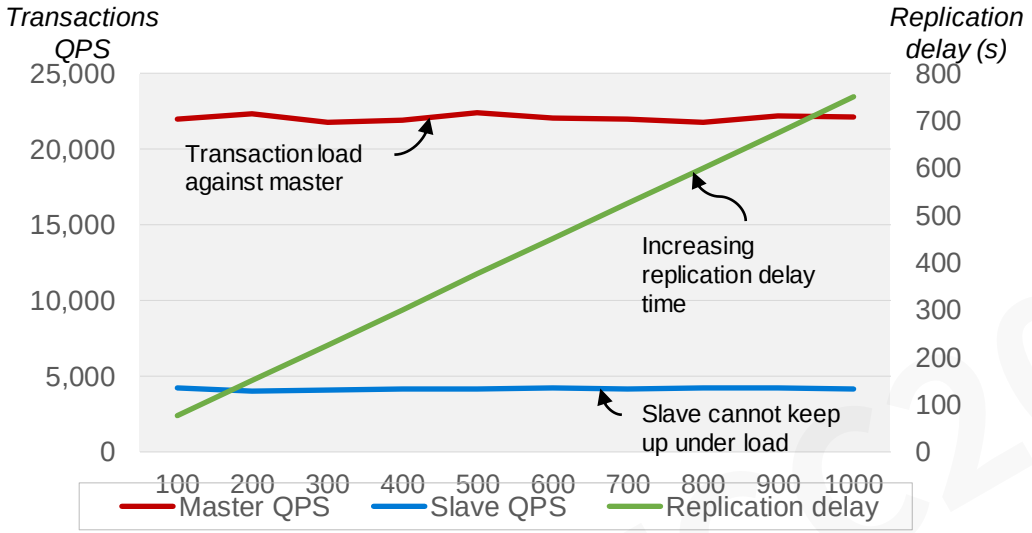
HWSQL

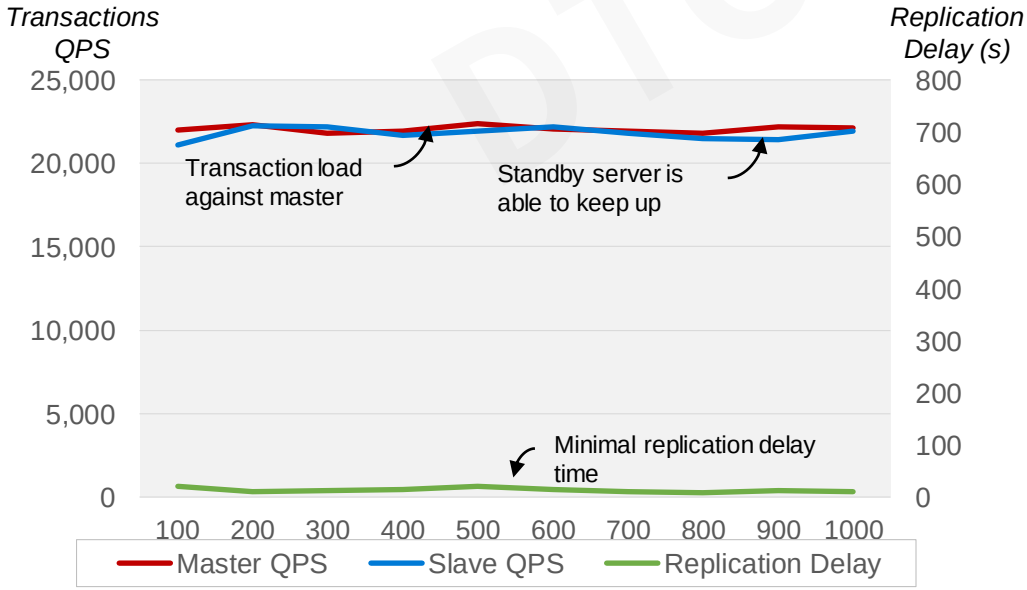


Thread-pool contains multiple thread groups (TG)

- Each TG can have multiple, reusable threads (e.g., 4096), but only 1 (or very few) active
- Connection is assigned to one of the TGs
- Statements from a connection can be served by different threads in its assigned TG
- Designed to distribute connections across TGs; no single TG becomes the bottleneck

Transaction-level MTS Parallel Replication





Logical clock based MTS

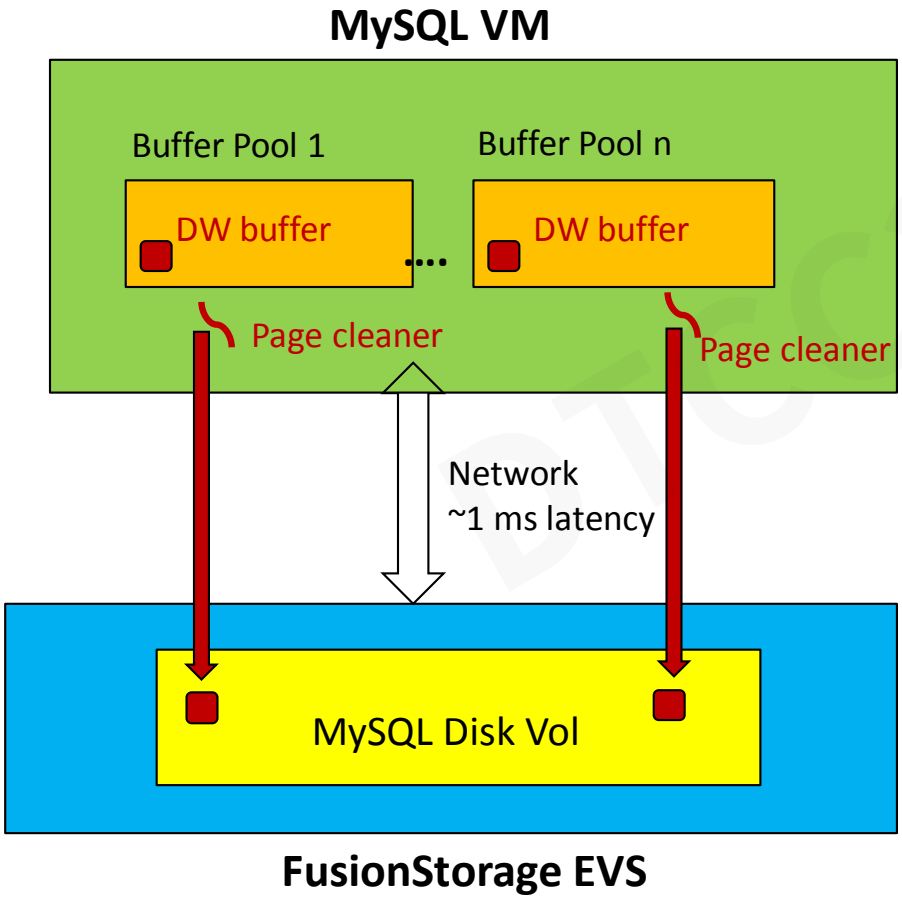
- Transactions are applied in parallel with multiple worker threads on slave if they do not interfere with each other, based on a scheduling algorithm of Logical Time Interval
- The parallelization is at transaction level (No matter which database the transactions are applied to)
- The slave SQL thread reads out transaction events from relay log, and schedules the transactions

Loss-less Semi-sync Replication

- Master waits for slave's ACK before committing (as opposed to: master waits for slave's ACK after committing).
  - Therefore, concurrent transactions do not see changes while this transaction waits for ack.
- Should a master fail, then any transaction that it may have externalized is also persisted on a slave.
- Master can optionally wait for multiple ACKs
  - Master does not commit transaction until it receives N ACKs from N slaves.

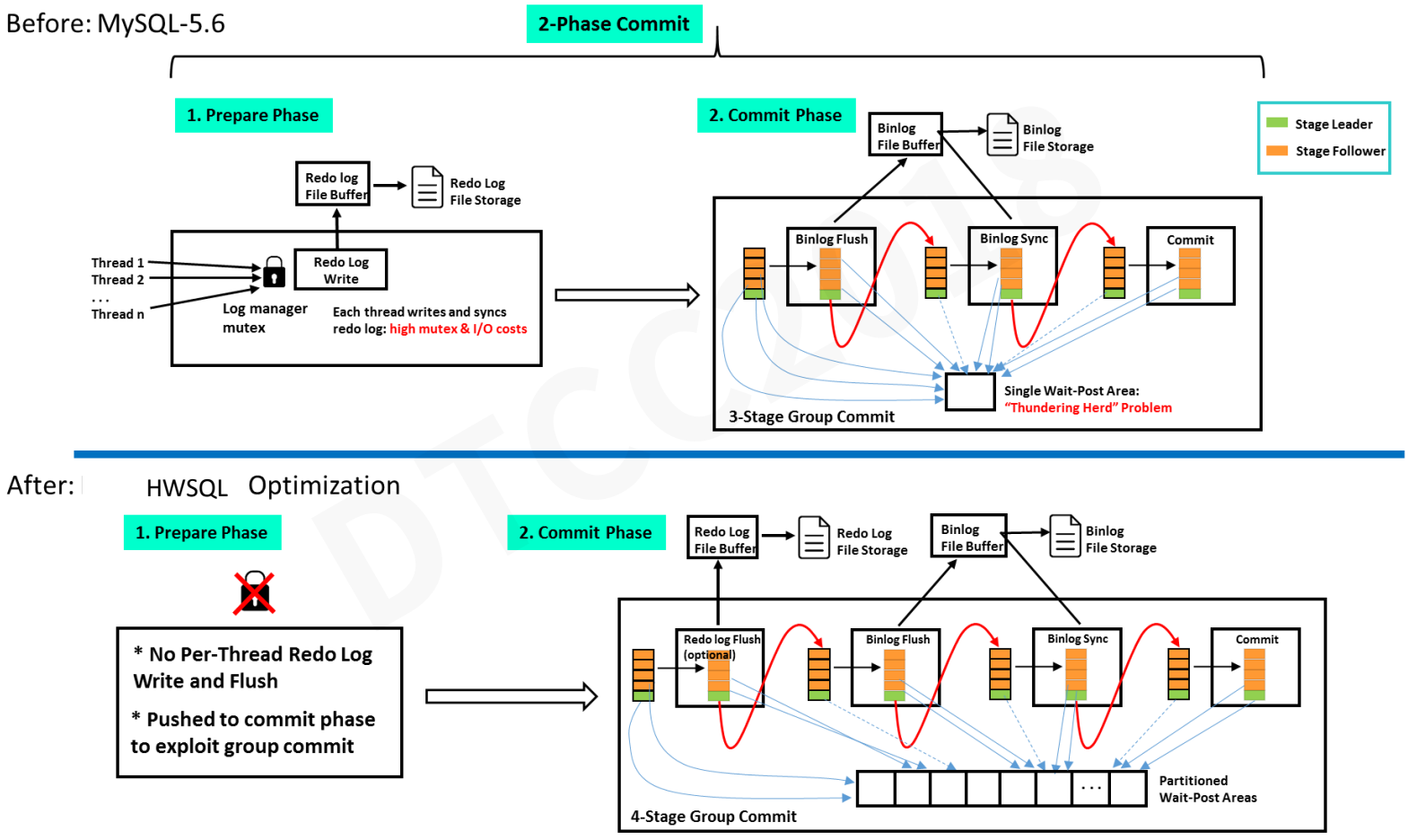
Page Write Optimizations

Average 1ms latency when MySQL VM visits FusionStorage EVS, severely impact the buffer pool write performance.



- HWSQL 5.6 implemented the parallel double write page flushing
- Changed from single double write buffer to one double write buffer per buffer pool instance。
- Introduced multiple page_cleaner threads.
- Ported adaptive page flushing algorithm

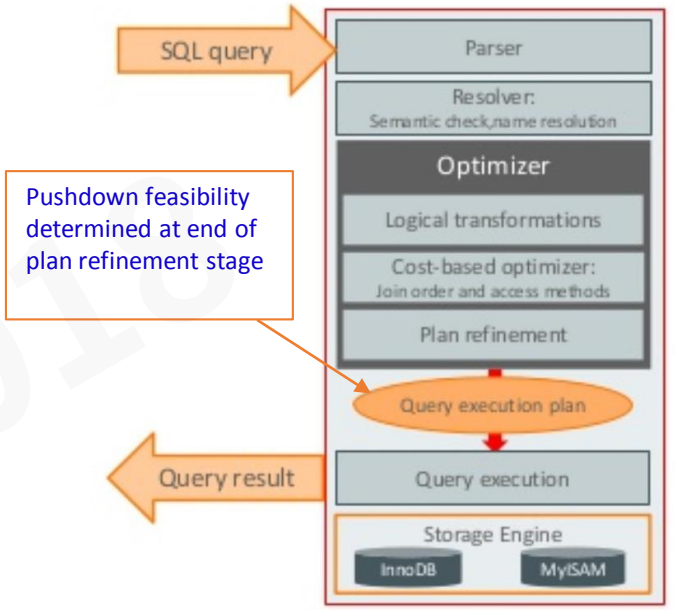
Group Commit & Notification Optimizations



SQL Aggregate Pushdown

HUAWEI

- Pushdown query aggregate evaluation to *storage engine* to reduce overheads between SQL server & storage engine
- Let the optimizer determine the best access plan, including which indexes are used to access the tables
- Can pushdown all aggregate functions (sum, count, etc) through common interface
  - Query group-by pushdown
  - Query condition evaluation pushdown (*eg, select count(*) from t1 where c1 > 3*)
- No API change and no compatibility impact
- Subquery, stored procedure, UDF not pushdownable

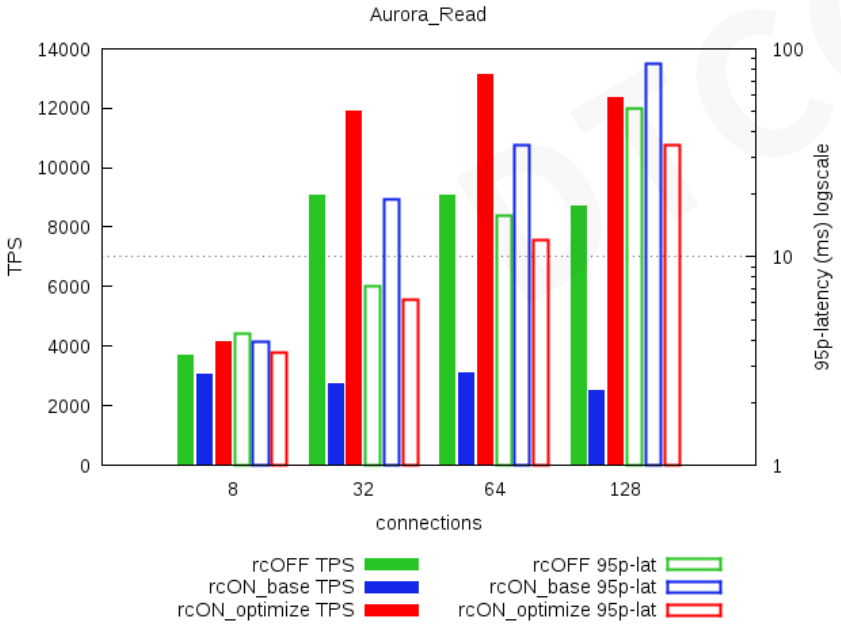


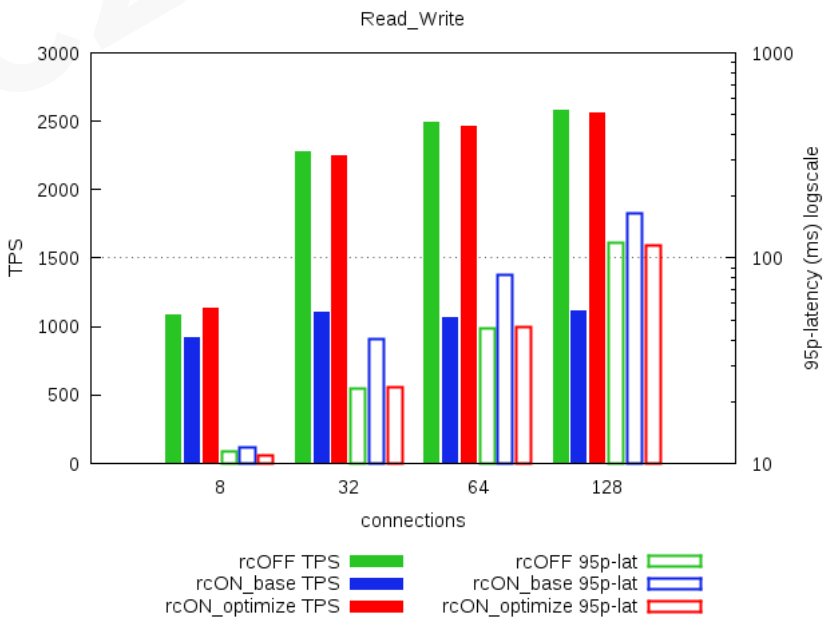
Example 1:
SELECT COUNT(*) FROM LINEITEM
lineitem is a table in a 10-GB TPC-H database. (roughly 60 million rows)
time reduced from 9.33s to 4.25s, i.e. 54%

Example 2:
select sum(k) from … where id between <range>
time reduced from 21.12s to 15.76s, i.e. 25%

Query Cache

- Partition global query cache into multiple segments
- Auto query cache deactivation
  - Global deactivation
  - Per-table deactivation
- Lock free structures and more efficient hashing





HUAWEI

Challenges & Opportunities

Main Challenges

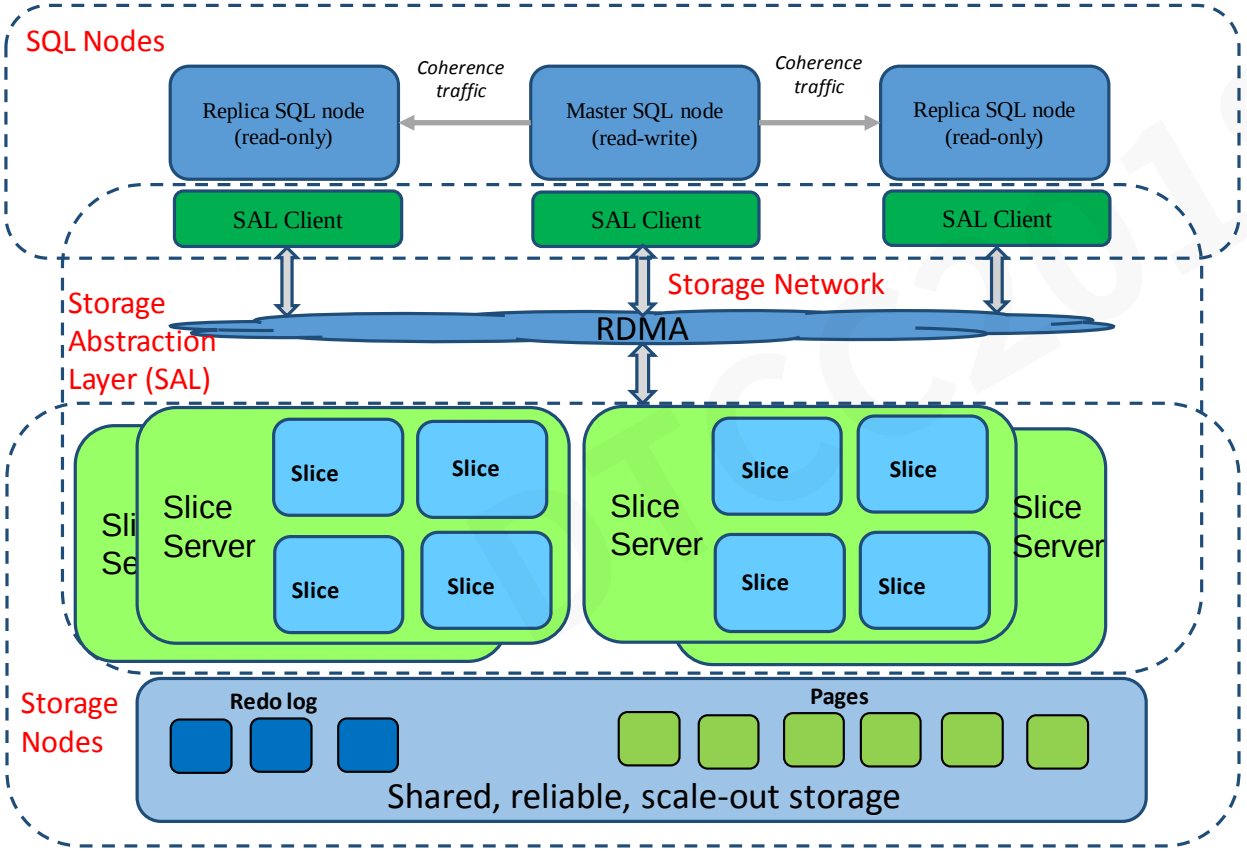
高安全性,高可靠性、高可用性、高性能，可扩展能力,以及运维自动化

- Traditional RDBMS architectures are 30+ years old
  - 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
- Auto-scaling, self-tuning

Huawei Cloud Native Database

- 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
  - Log is the database, 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

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

THANKS


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