Alibaba Group
阿里巴巴集团

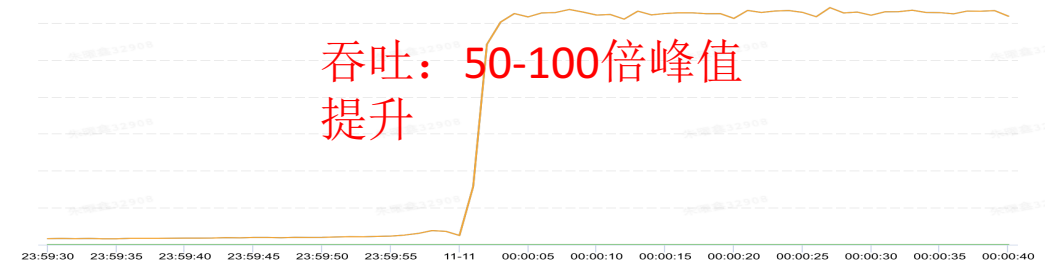
阿里新一代分布式数据库X-DB架构设计与技术剖析

曲山
阿里巴巴资深技术专家
2018-10-18

阿里业务对数据库带来的挑战

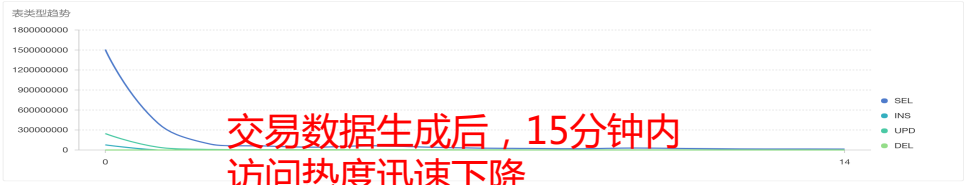
阿里业务对数据库的挑战-性能

- **双11购物节**
  - 性能、稳定性的双重考验
  - 核心业务
    - 交易、库存、购物车、优惠…
  - 瞬时冲击
    - 50-100倍瞬时峰值的流量冲击
  - 响应延时
    - RT基本保持稳定，确保用户的体验不受影响

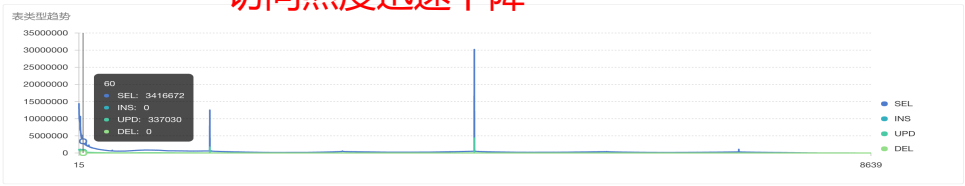


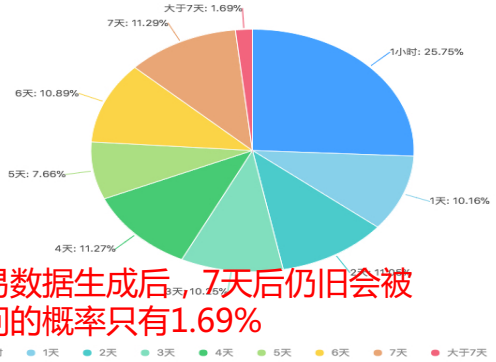
阿里业务对数据库的挑战-成本

- **数据量大**
  - 存储空间大、成本高
- **数据冷热分离特性明显**
  - 如何基于数据的冷热特性，提升整体数据库的存储效率
- **交易数据分级存储现状**
  - 历史数据库和在线数据库分离，定期迁移
  - 用户体验差，应用开发复杂



交易数据生成后，15分钟内
访问热度迅速下降

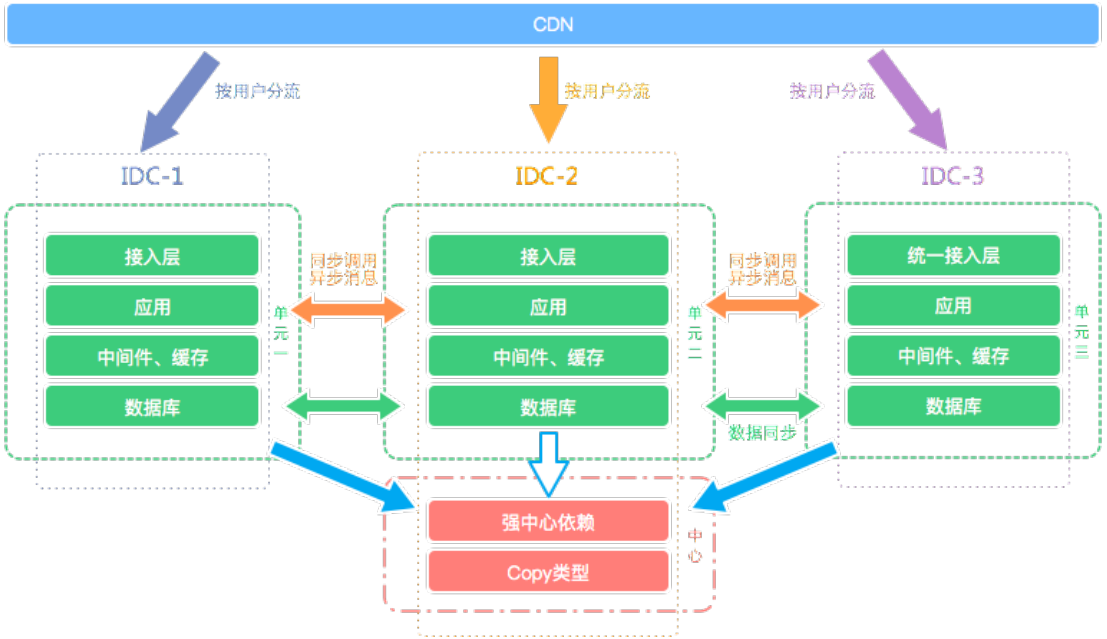




交易数据生成后，7天后仍旧会被
访问的概率只有1.69%

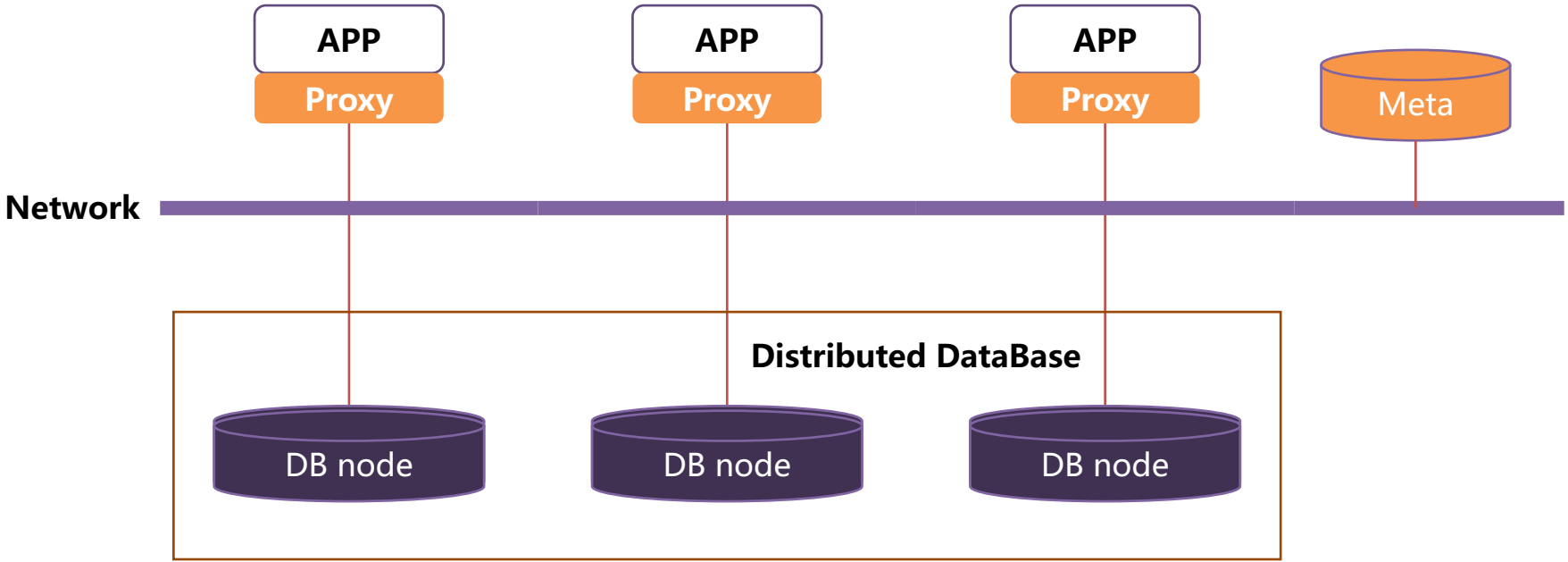
阿里业务对数据库的挑战-跨域高可用

- **异地多活：跨域高可用**
  - 传统银行解决方案：两地三中心
  - 阿里巴巴解决方案：异地多活
- **异地多活最大的考验**
  - 数据库集群跨域部署
  - 数据库单机、AZ、Region级别持续可用，对应用透明
- **全球化战略**
  - 异地多活 -> 全球化部署

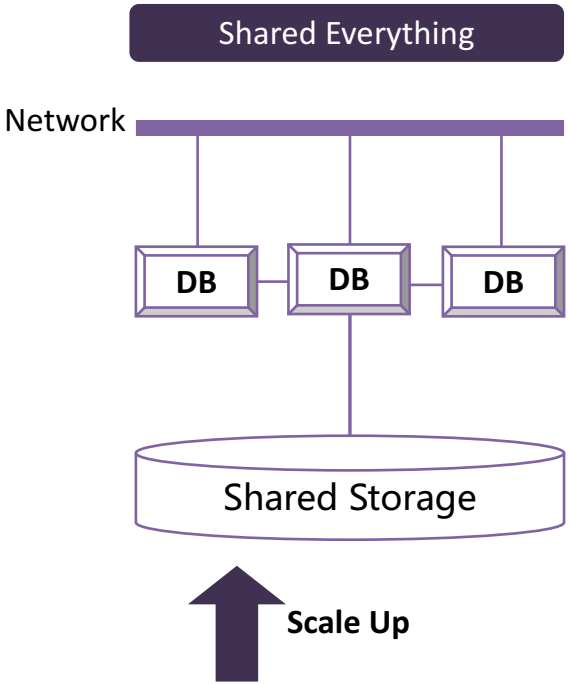


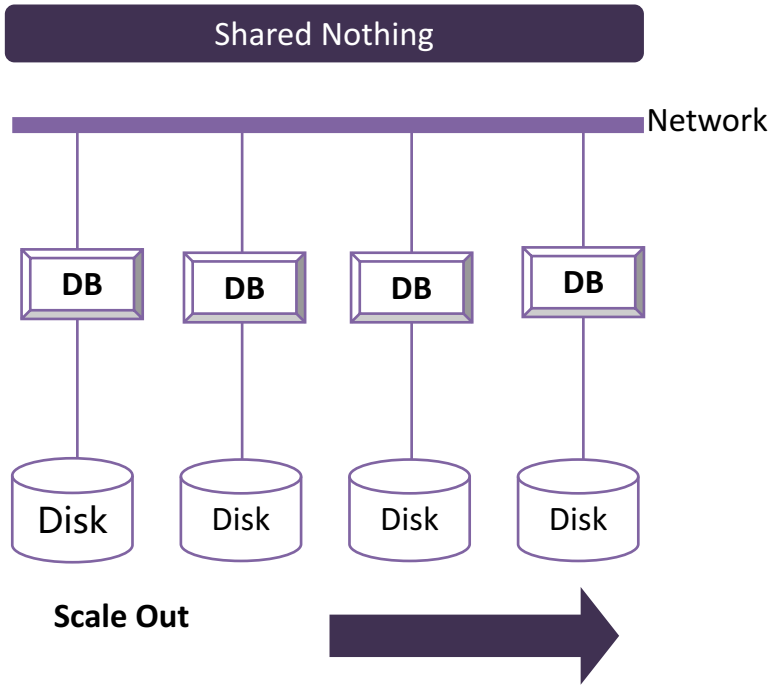
阿里数据库架构的演进

单机数据库到分布式数据库



shared-storage & shared-nothing





X-DB 架构设计

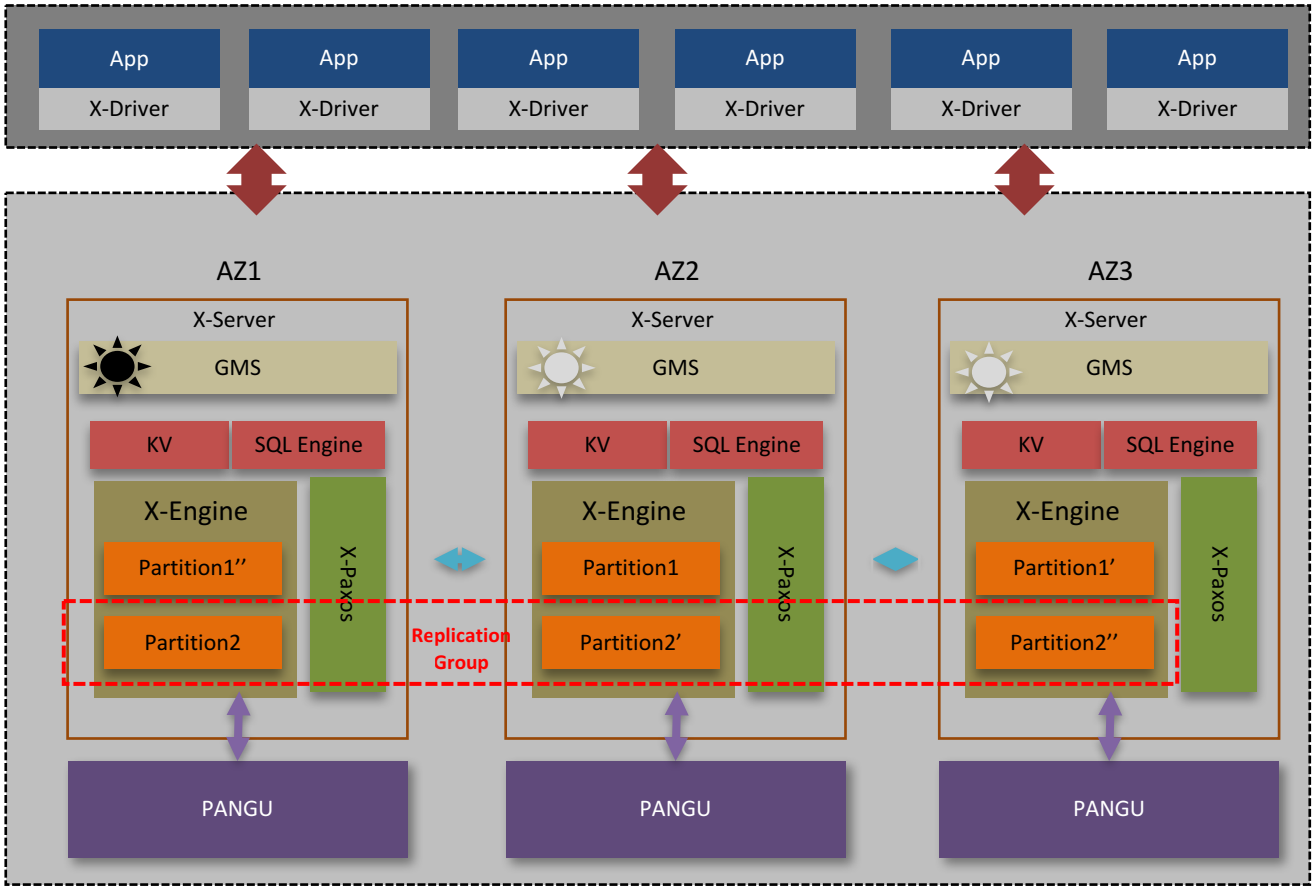
X-DB : shared-nothing architecture

X-DB

- Shared-Nothing
- Globally Geo-Distributed
- Strong ACID Guarantees
- Layered Storage
- Horizontally scaling

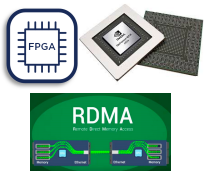
Key Components

- GMS
- SQL and KV Engine
- X-Engine
- X-Paxos

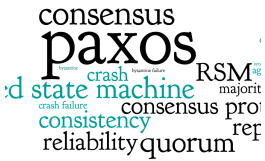


X-DB 核心组件与技术

X-Engine: 基于冷热数据分离的分层存储引擎，高性能（目标1M TPS)，低成本



基于软硬件结合的设计思想，充分发挥FPGA/GPU等异构计算设备的强大算力，以及RDMA/NVM等新型网络及存储硬件的时延优势。



拥有全球级部署能力，多种一致性级别，自适应行级多点写入的分布式一致性协议

GMS：存储和计算的分布式弹性扩展与负载均衡，基于混合时间戳的轻量级、高性能分布式事务

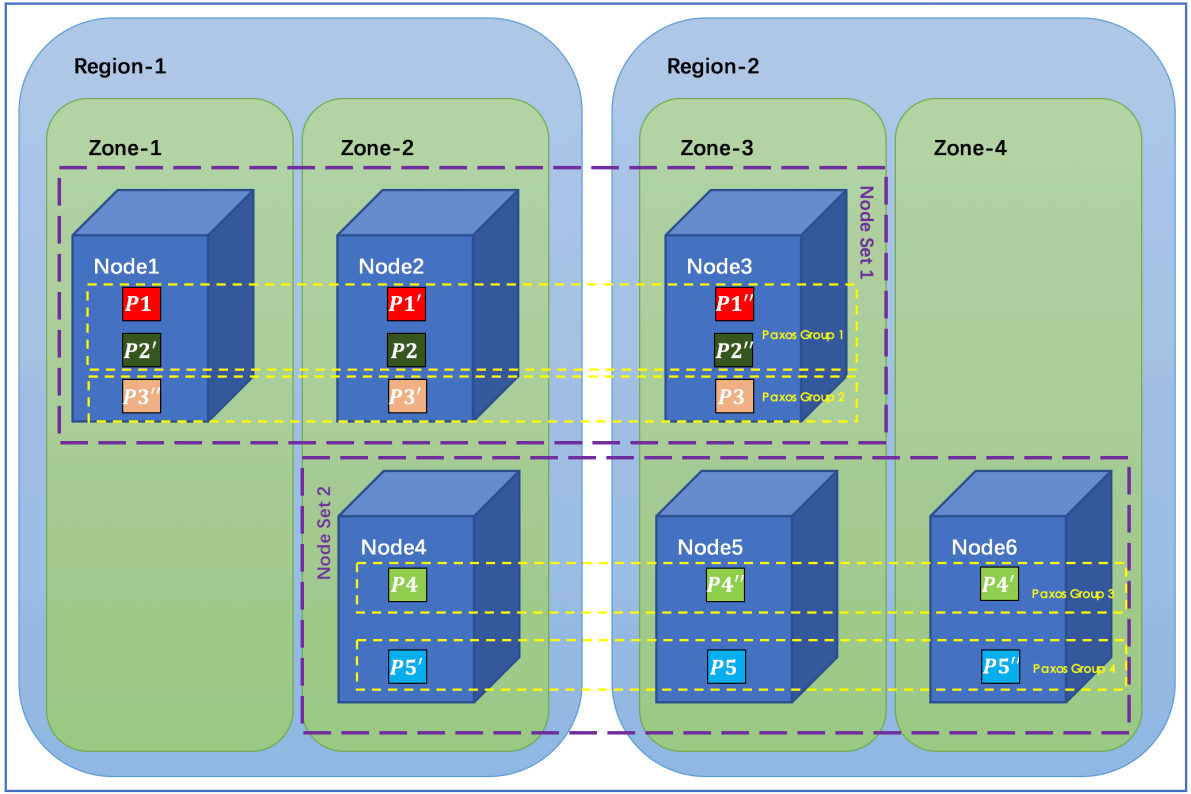
SQL Engine: 一体化高性能分布式SQL处理引擎，实现独特的基于LSM-Tree存储的优化器模型，支持跨节点复杂查询和一致性读。

X-DB技术剖析

data sharding & multi-master

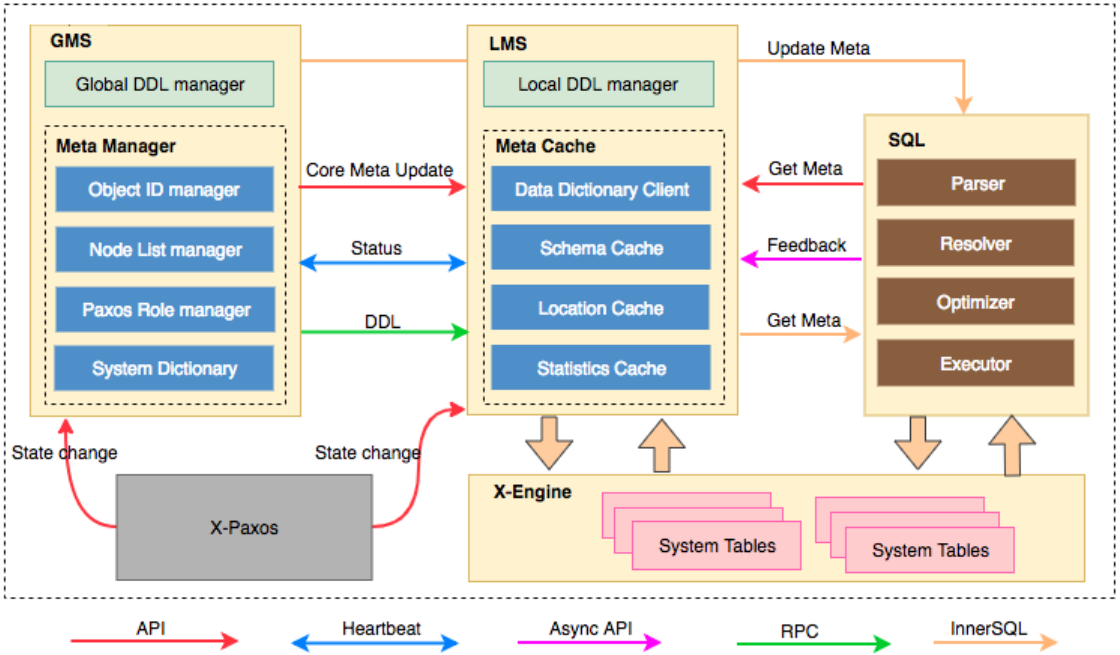
data sharding

- partition by hash(range)
- multi paxos group
- multi master



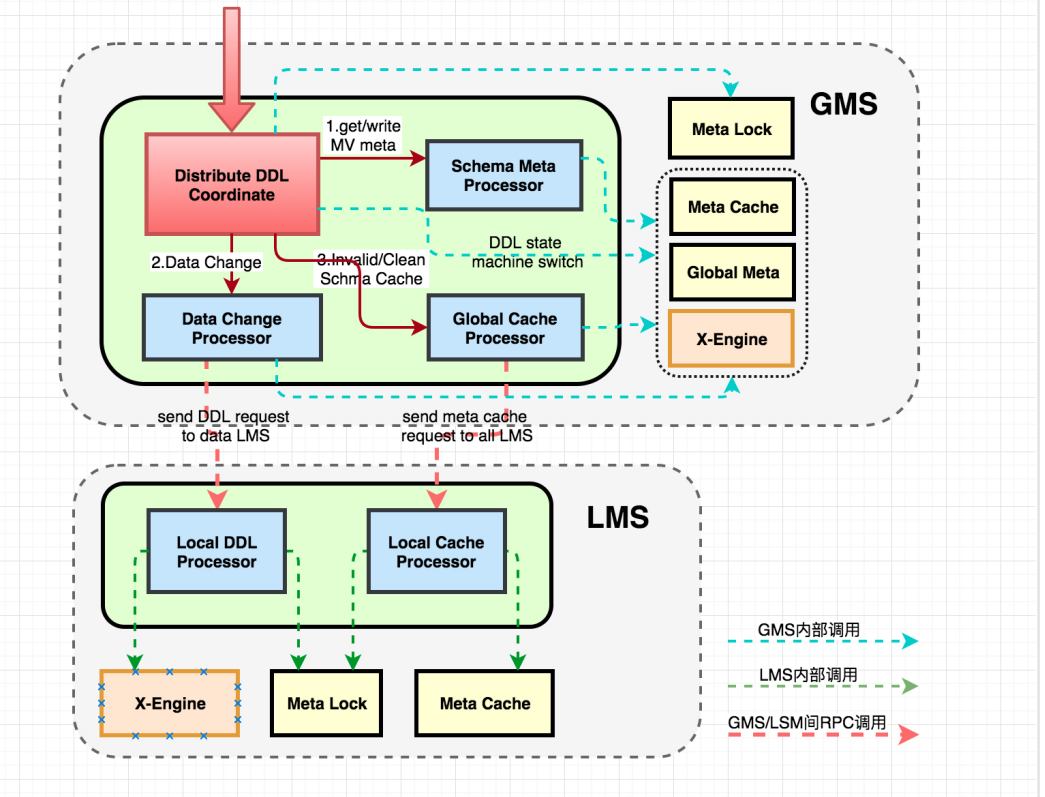
meta data management

- 全局只有一份元数据
- 元数据持久化
- 元数据多级缓存



distribute DDL

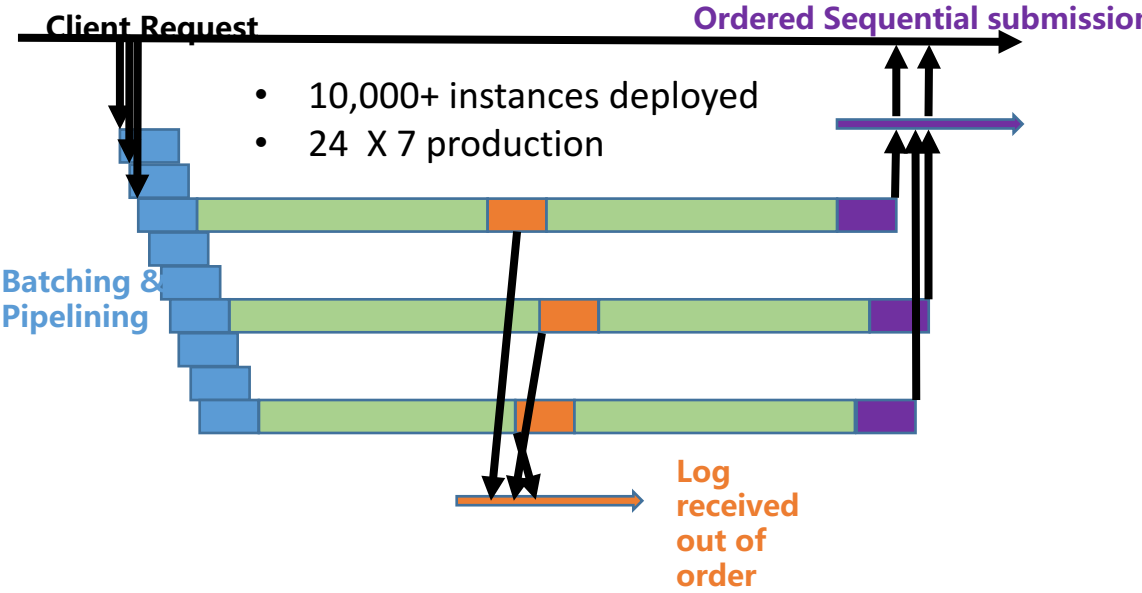
- Online DDL
  - Schema变更不阻塞DML
  - 多版本Schema
  - 理论上所有的Schema变更都可以支持Online
- Fast DDL
  - 只修改元数据、不触碰用户数据，需要X-Engine支持
  - Schema下沉
  - 也属于Online DDL

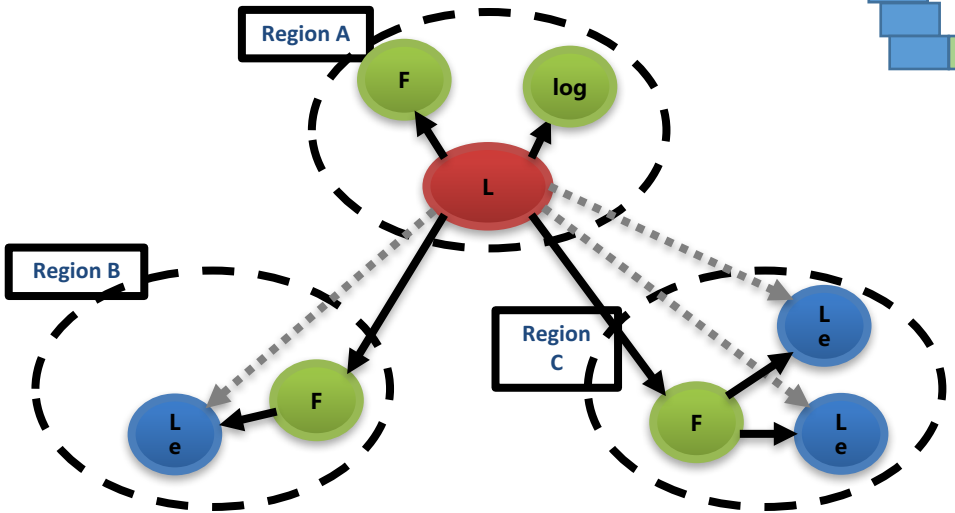


X-Paxos: GEO-distributed & HA & strongly consisctency

X-Paxos

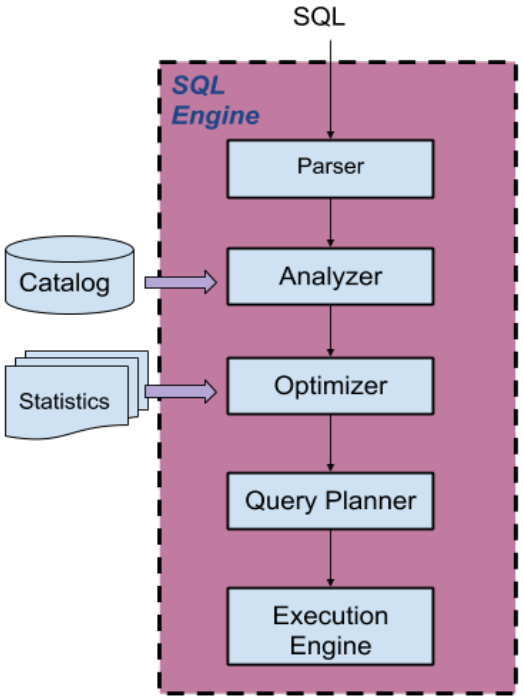
- Cross-Region HA
- Global Replication
- Flexible Topology
- Interchangeable Roles



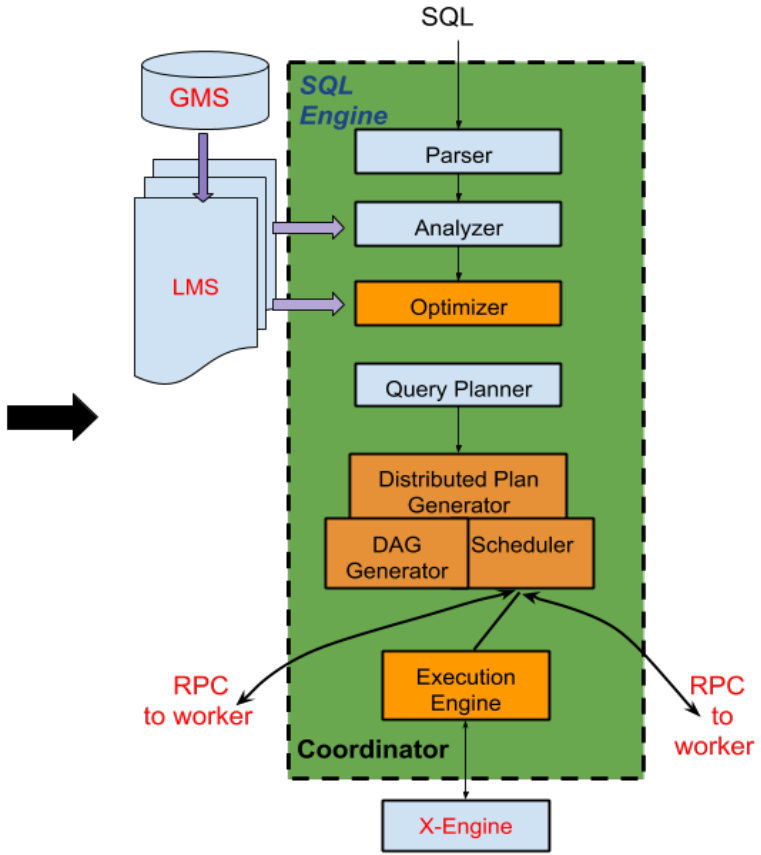


SQL Engine

- 单机(优化器)
- 分布式执行引擎

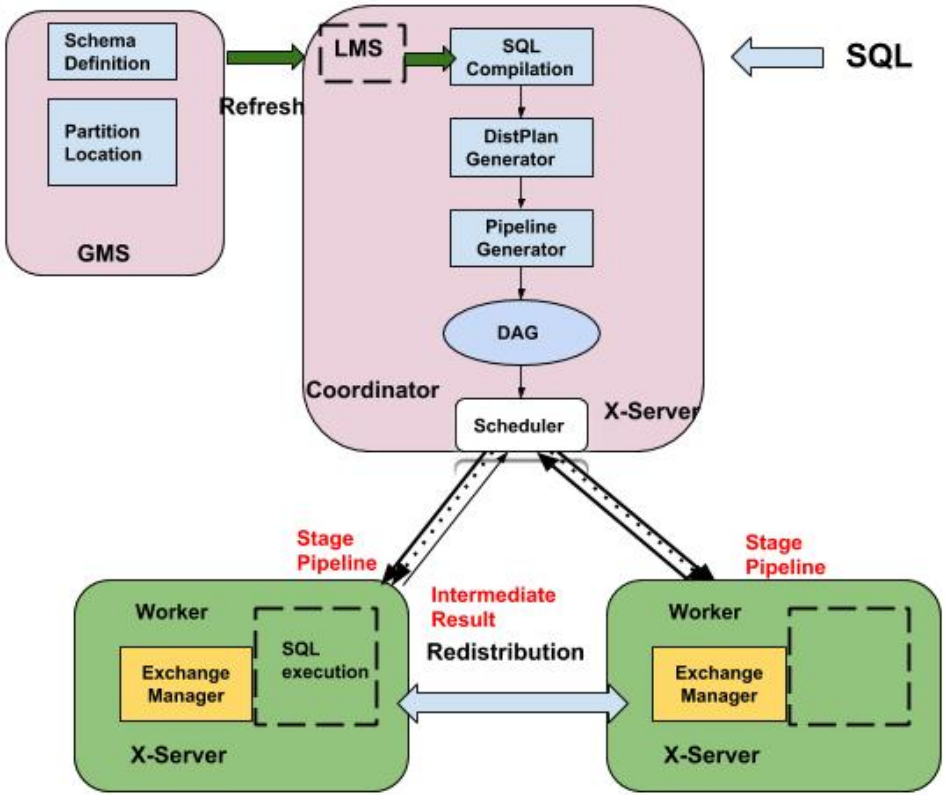


单机



分布式(X-DB)

SQL Engine (distribute execution)



Distributed Architecture

- Massively Parallel Processing
- DAG based plan with scheduling
- Pipeline execution

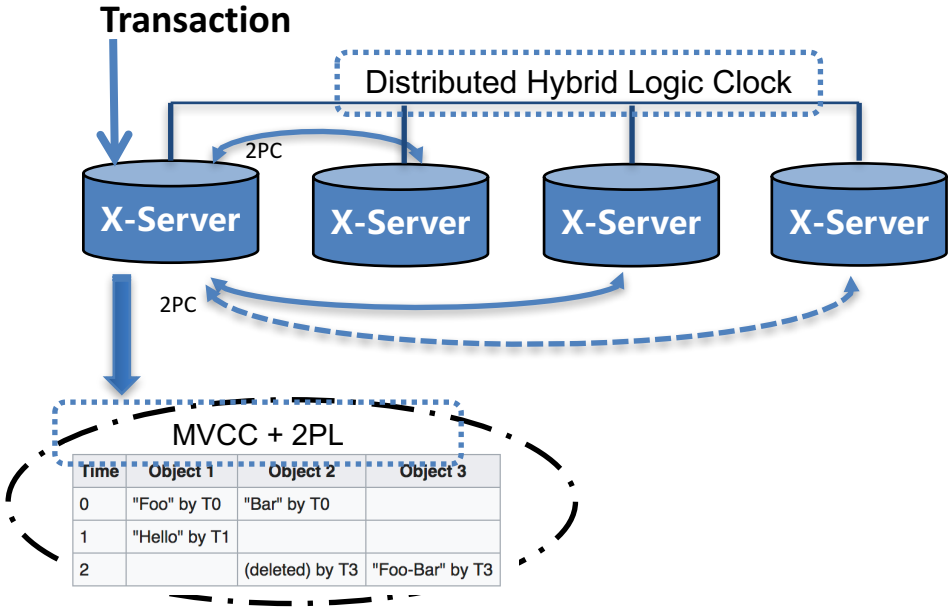
Execution Engine

- Resumable piece-meal Execution
- Exchange Manager

distribute transaction

Distributed Transaction Design Principle

- Great majority of transactions only touch single shard(aka. partition)
- Scalability is essential
- Provides **A**tomicity + **I**solation in ACID
- Decentralized to remove SPoF

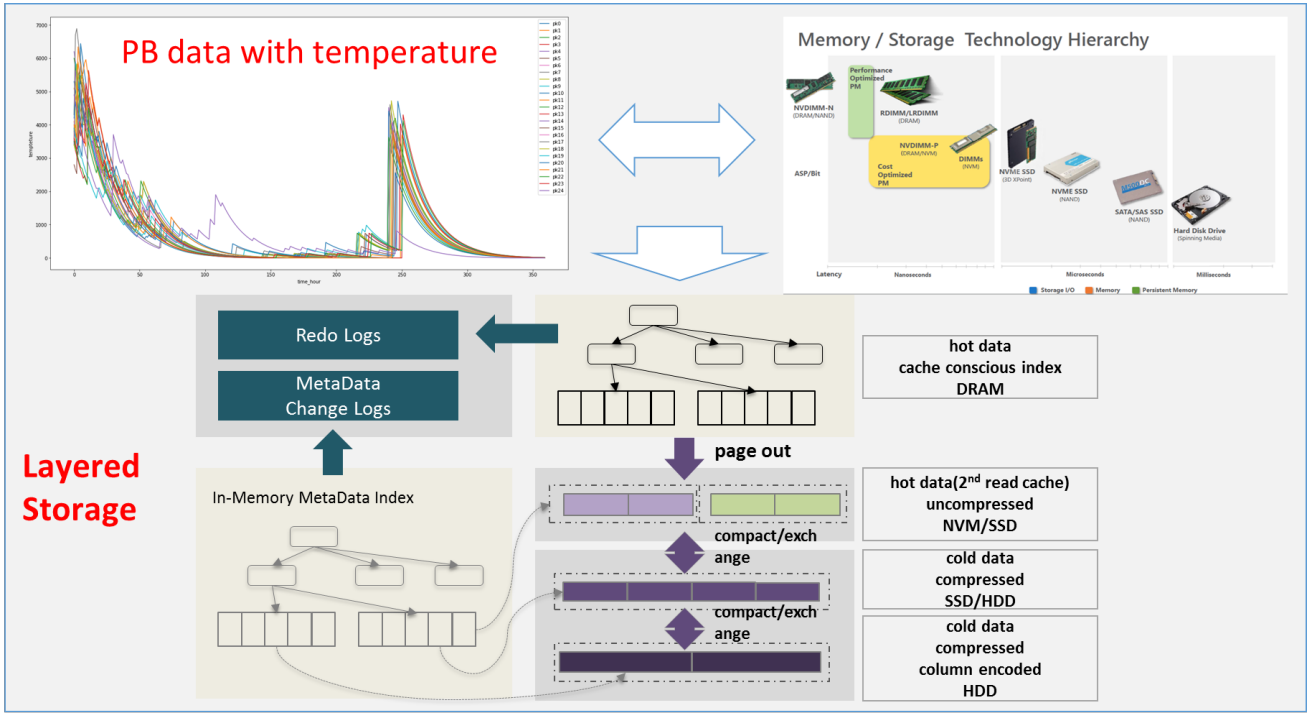


Isolation Level: SSI, **SI, RC**; Concurrency Control: **MVCC+Lock**; Snapshot Timestamp: **Hybrid Logical Clock**

Daniel Abadi. NewSQL database systems are failing to guarantee consistency, and I blame Spanner

Tiered Storage Engine

- **Huge Data Volume:** High demand on storage capacity
- **Hot-Cold Data :** Access and Storage efficiency



How to optimize LSM-like system

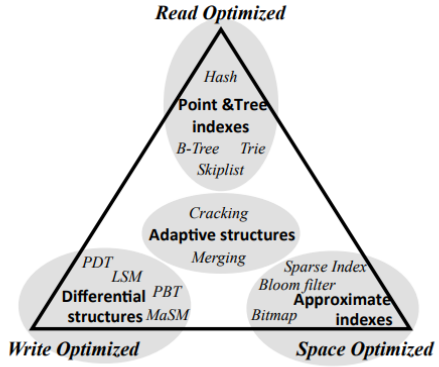


Figure 1: Popular data structures in the RUM space.

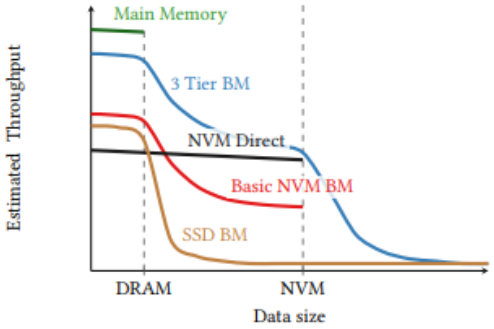
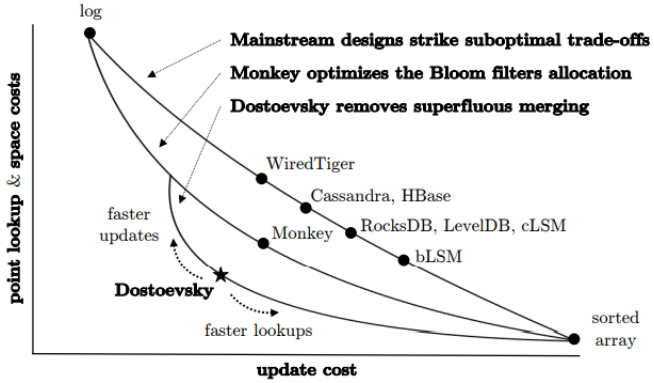
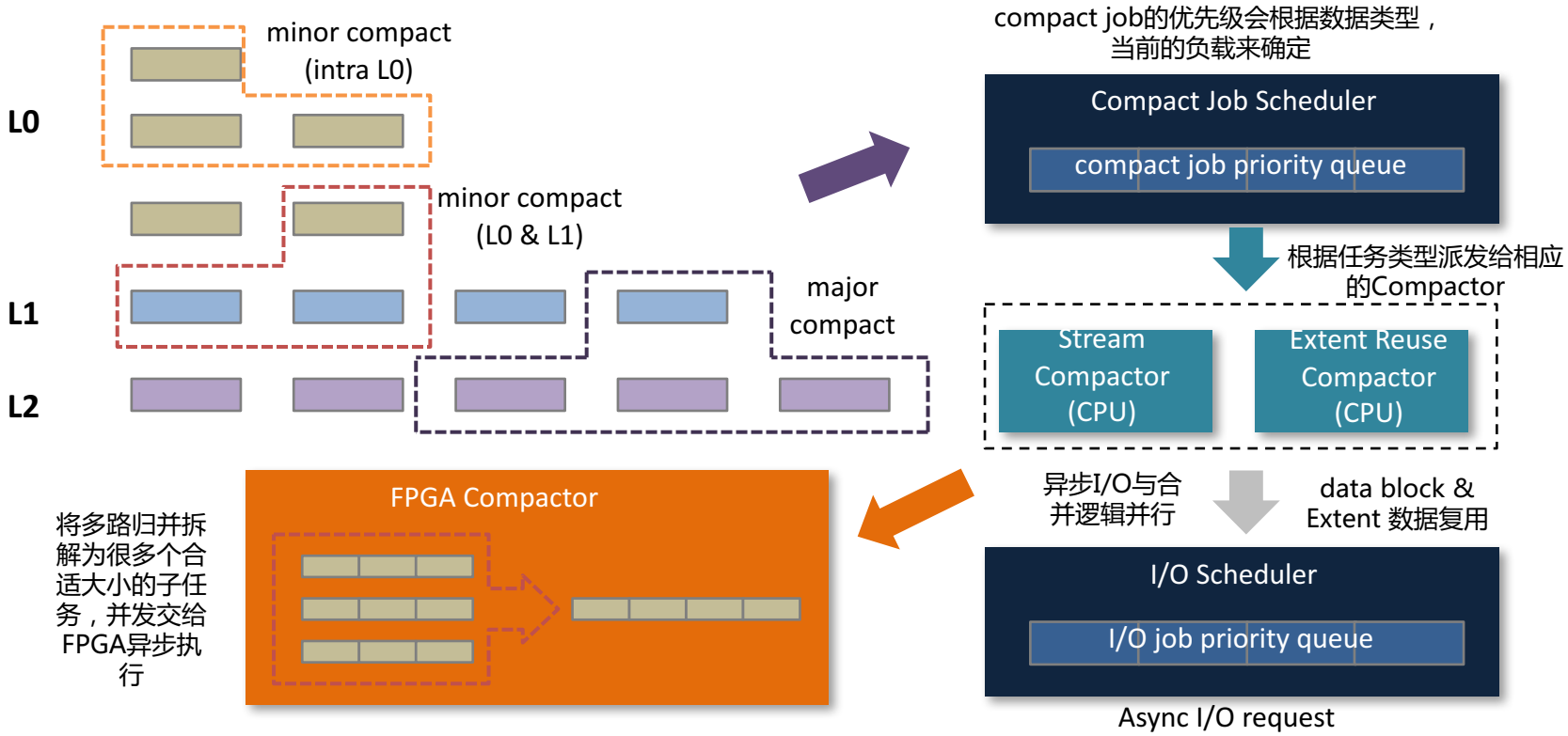


Figure 1: System designs under varying data sizes.



- Alexander van Renen. 2018. Managing Non-Volatile Memory in Database Systems
- Huanchen Zhang. 2018. SuRF: Practical Range Query Filtering with Fast Succinct Tries (SIGMOD 2018 Best Paper Award)
- Niv Dayan. 2018. Dostoevsky: Better Space-Time Trade-Offs for LSM-Tree Based Key-Value Stores via Adaptive Removal of Superfluous Merging (Harvard Data Systems Lab)
- Ildar Absalyamov. 2018. Lightweight Cardinality Estimation in LSM-based Systems

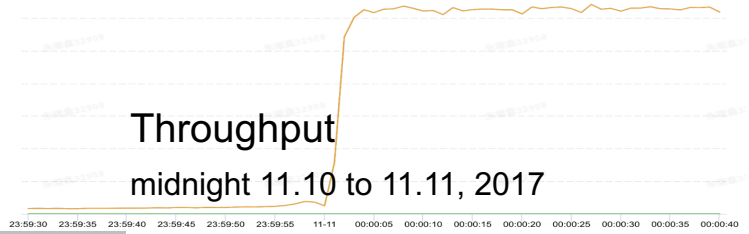
Hardware Acceleration

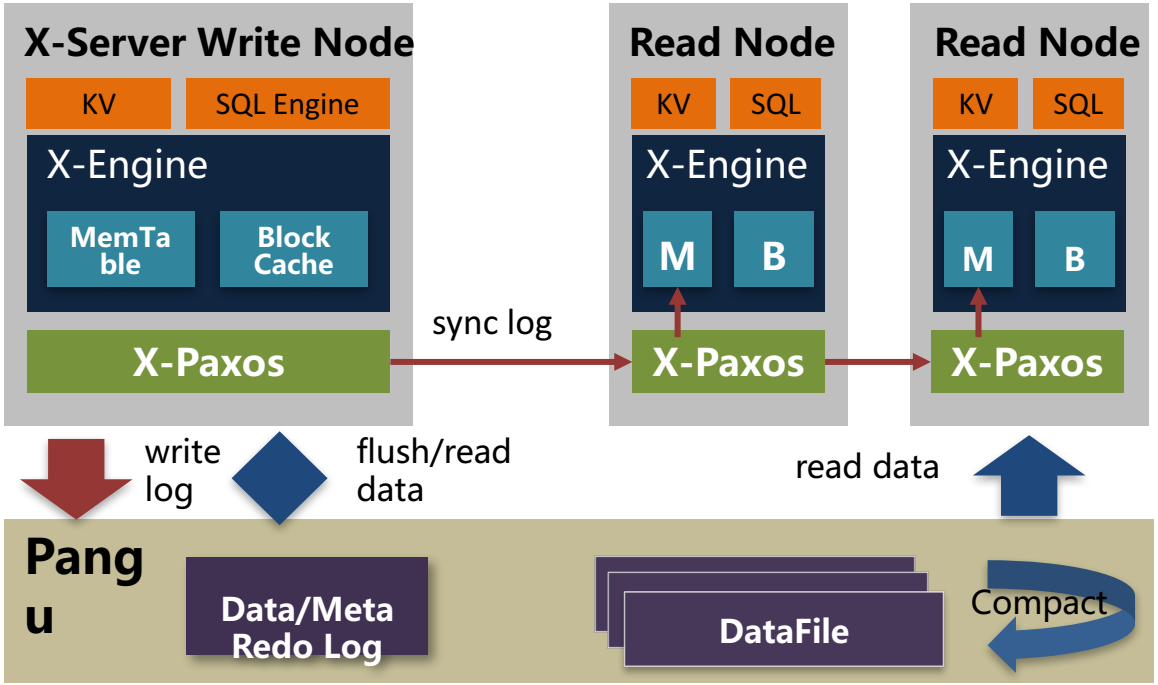


decouple computation and storage

Singles' Day Shopping Festival

- Throughput: 50x~100x than ordinary
- Response Time: fast and stable <0.9ms
- Cost：requires elastic scaling





Decoupling: High Elasticity

- Stateless computing nodes
- X-Engine push down computation
- Distributed FS Pangu(盘古)
- Simple and flexible cloud deployment with container technology

Summary: Challenge and Opportunity

- HW-SW co-design
  - ✓ **Multi-Core/NUMA:** Parallel logging
  - ✓ **FPGA:** Data compaction for storage engine & Compression, SQL evaluation, JOIN …
  - ✓ **NVM:** Multi-tier storage for hot/cold data & Learned Index
  - ✓ **RDMA:** Logging for Paxos
  - ✓ **GPU:** Workload prediction & Hot/cold data prediction
- SQL and Optimizer
  - ✓ **Parallel** Execution
  - ✓ **LSM-tree Based** Cost Model
  - ✓ **Distributed Execution**
- Storage engine
  - ✓ Decouple computation and storage engines for high **elasticity**
  - ✓ High **throughput** low **latency** under high concurrency
  - ✓ **Hot-cold** data management
- AI-Based Intelligent Database
  - ✓ **CloudDBA**: Smarter and Faster than DBA
  - ✓ **Self-**Diagnose and Fix
  - ✓ Intelligent **Data Processing**
  - ✓ Predict Future **Trends** and Adapt to **Changes**
  - ✓ **ML-Based** optimizer

Thanks & Questions

欢迎加入!