Alibaba Group
阿里巴巴集团

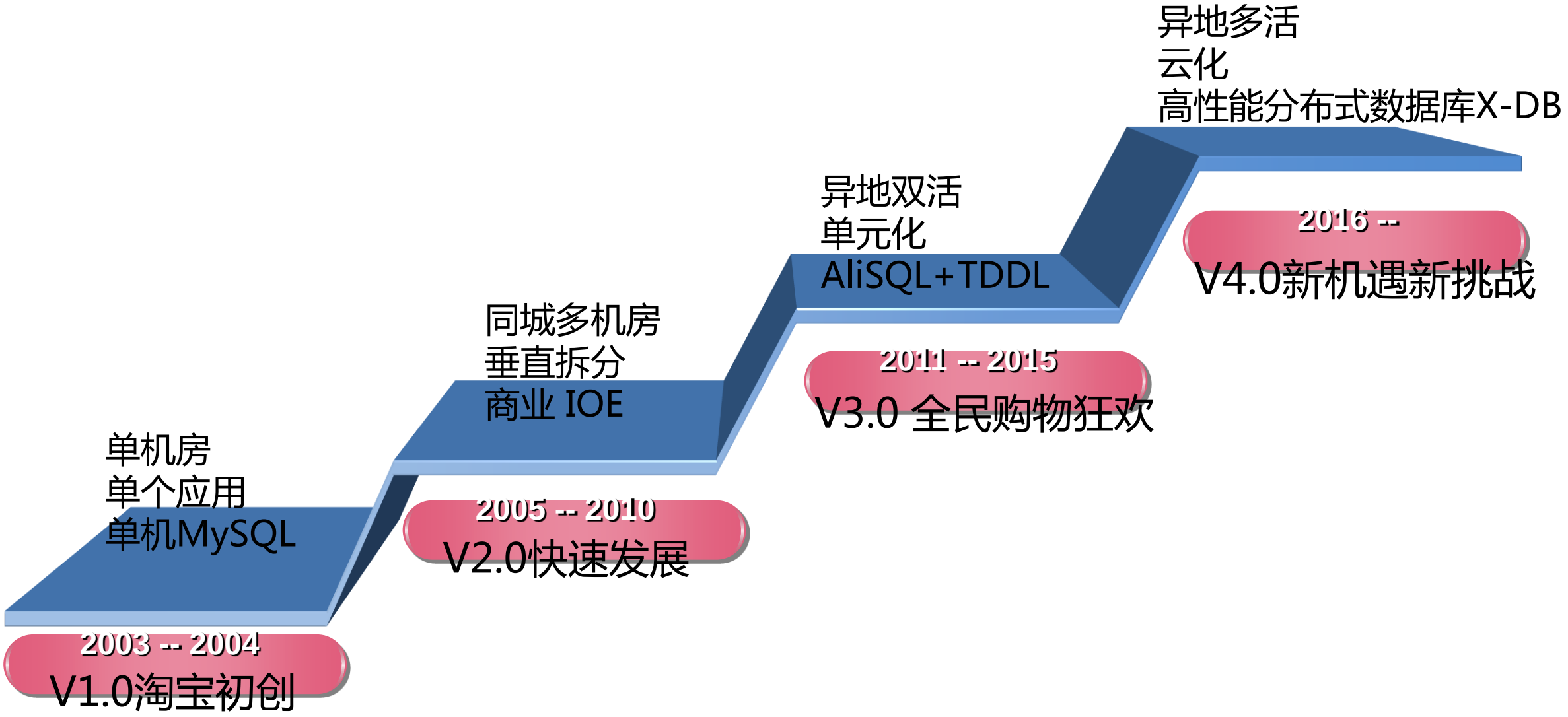
Database in Alibaba：Bridge between Theory and Practice

阿里巴巴-数据库事业部-何登成

何登成，花名圭多。阿里巴巴资深技术专家，阿里巴巴数据库内核研发团队负责人。

- 浙江大学计算机学院本科、研究生，师从陈刚老师。2005年至今，一直专注在数据库领域，先后在神州通用、网易、阿里从事数据库研发和管理工作
- 连续多年阿里巴巴双11、双12、支付宝新春红包大型活动数据库总负责人
- 目前带领阿里巴巴数据库内核研发团队，打造下一代集分布式、持续可用、高性能、低成本于一体的关系型数据库系统

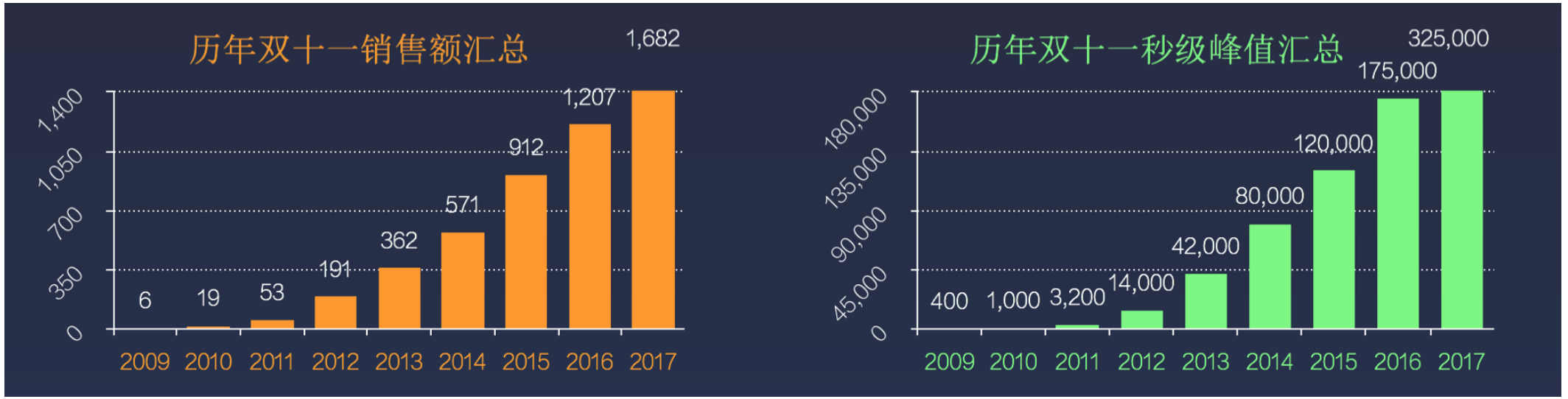
阿里巴巴数据库编年史


单机房
单个应用
单机MySQL
2003 -- 2004
V1.0淘宝初创
同城多机房
垂直拆分
商业 IOE
2005 -- 2010
V2.0快速发展
异地双活
单元化
AliSQL+TDDL
2011 -- 2015
V3.0 全民购物狂欢
异地多活
云化
高性能分布式数据库X-DB
2016 --
V4.0新机遇新挑战

大纲

- 阿里巴巴OLTP数据库应用特点
- 业界数据库发展的关键里程碑
- X-DB：阿里自研高性能分布式数据库
- 阿里巴巴数据库：研究和挑战

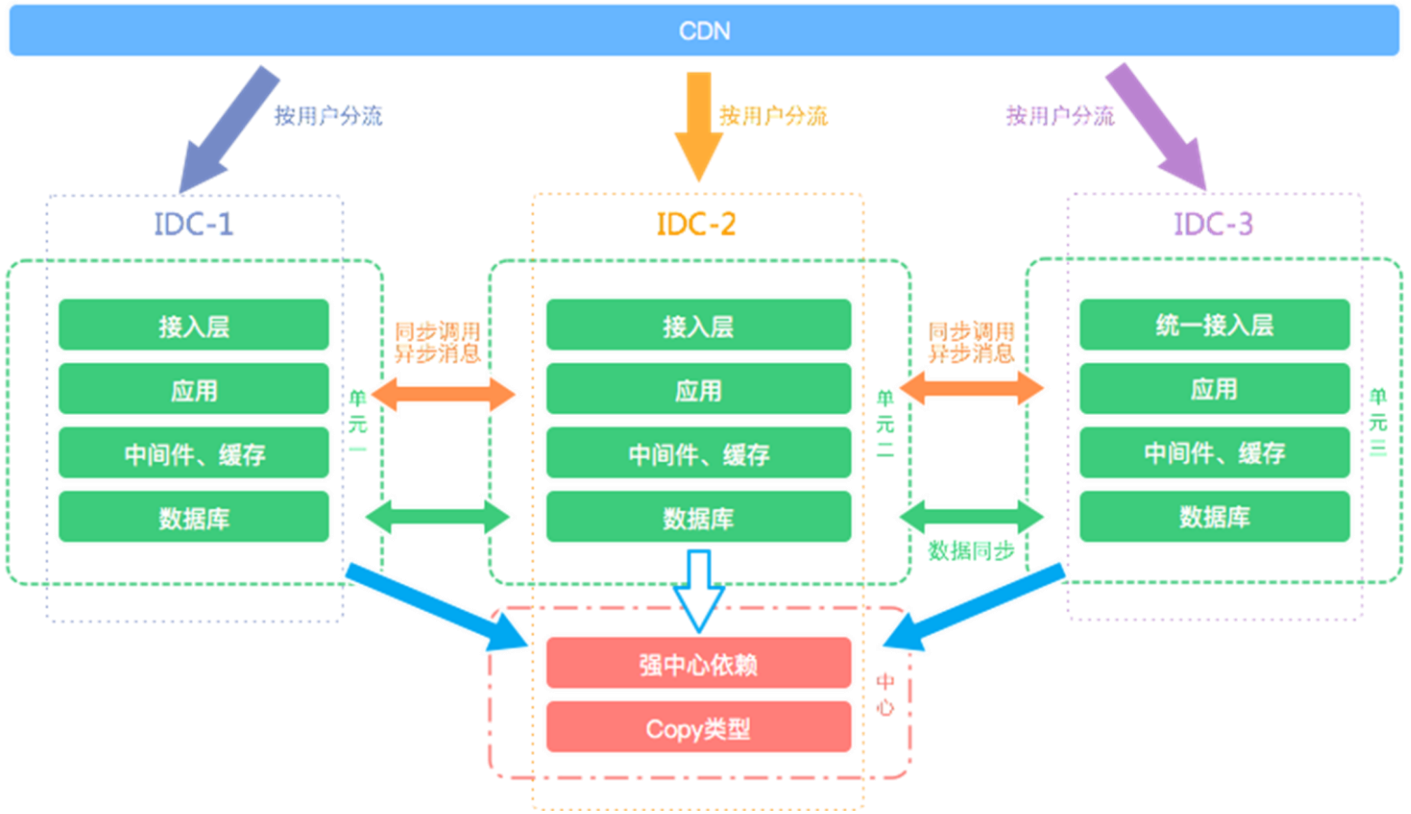
双11：一场技术大练兵



- 高性能：支撑尽可能高的零点峰值，给用户最好的体验
- 低成本：成本要尽可能低，要求极致的弹性能力
- 稳定：系统性工程，互联网界的超级工程

异地多活：数据库面临的最大挑战

- **异地多活：高可用**
  - ✓ 传统银行解决方案：两地三中心
  - ✓ 阿里巴巴的解决方案：异地多活
- **异地多活最大的考验来自数据库**
  - ✓ 数据库集群跨Region部署
  - ✓ 数据库单机、AZ、Region级别的持续可用，对应用透明
- **阿里巴巴全球化战略**
  - ✓ 异地多活 -> 全球化部署



存储：效率和成本间的综合考量

- **数据量大**
  - ✓ 数据存储空间要求

- **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

大纲

- 阿里巴巴OLTP数据库应用特点
- **业界数据库发展的关键里程碑**
- X-DB：阿里自研高性能分布式数据库
- 阿里巴巴数据库：研究和挑战

数据库发展关键里程碑

- **关系型数据库和SQL（**1970年后**）**
  - ✓ 诞生了一批数据库巨头
    - 商业：Oracle、微软SQLServer、IBM DB2
    - 开源：MySQL、PostgreSQL

- **分布式NoSQL**（2000年后，互联网兴起）
  - ✓ Google BigTable、Apache Hbase、Facebook Cassandra、Amazon DynamoDB、MongoDB

- **现代化数据库**（2010年后）
  - ✓ Google Spanner、微软 Hekaton、SAP HANA、慕尼黑工业大学 HyPer、Oracle 秘密项目（进行中）、VoltDB（Stonebraker创业产品）

2010年后，现代化数据库兴起（理论依据）

- OLTP Through the Looking Glass, and What We Found There
  - ✓ By Stavros Harizopoulos, Daniel J. Abadi, Samuel Madden, Michael Stonebraker
  - ✓ **2B3L**
    - B+Tree、Buffer Pool
    - Logging、Latching、Locking
- **数据库新技术奠基之作**
  - ✓ 微软：Hekaton
  - ✓ 慕尼黑工大：HyPer
  - ✓ Stonebraker：VoltDB

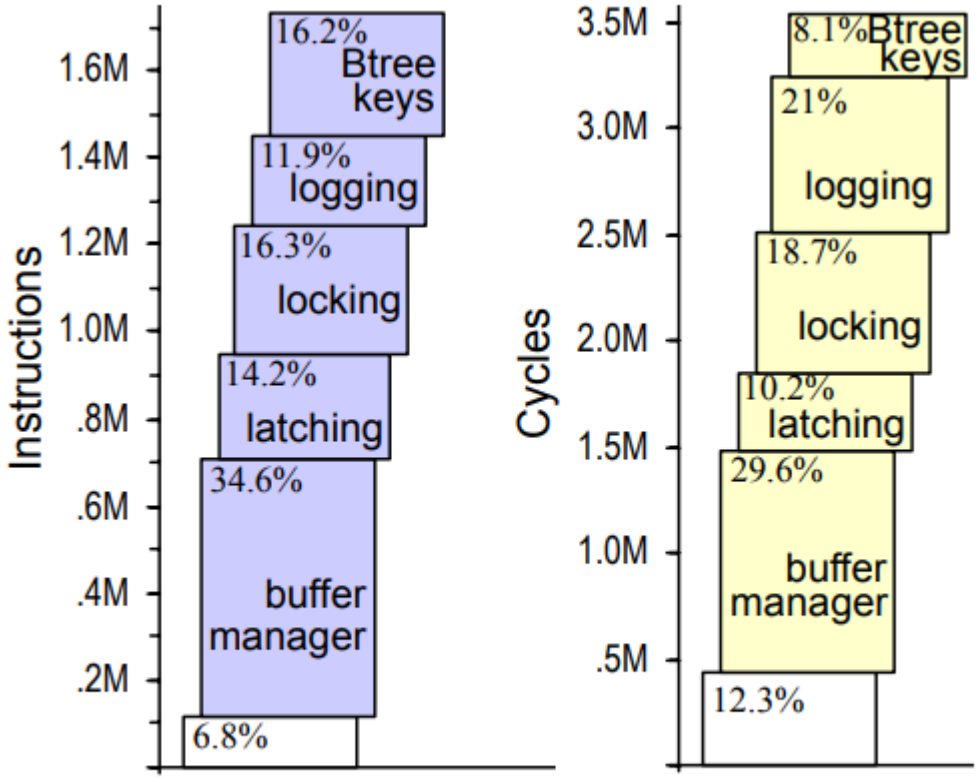


Figure 8. Instructions (left) vs. Cycles (right) for New Order.

大纲

- 阿里巴巴OLTP数据库应用特点
- 业界数据库发展的关键里程碑
- **X-DB：阿里自研高性能分布式数据库**
- 阿里巴巴数据库：研究和挑战

X-DB是什么？

- **阿里在线业务的需求**
  - ✓ 分布式
  - ✓ 高性能
  - ✓ 高可用
  - ✓ 高可扩展性
  - ✓ 低成本
  - ✓ 充分发挥新硬件效率
  - ✓ …

- **X-DB**
  - ✓ 阿里巴巴自研高性能分布式数据库
- **愿景**
  - ✓ 世界最快、成本最低的OLTP数据库
- **设计理念**
  - ✓ 关注用户使用效率，全面兼容MySQL生态
  - ✓ 关注软硬件Co-Desion，充分发挥硬件效率

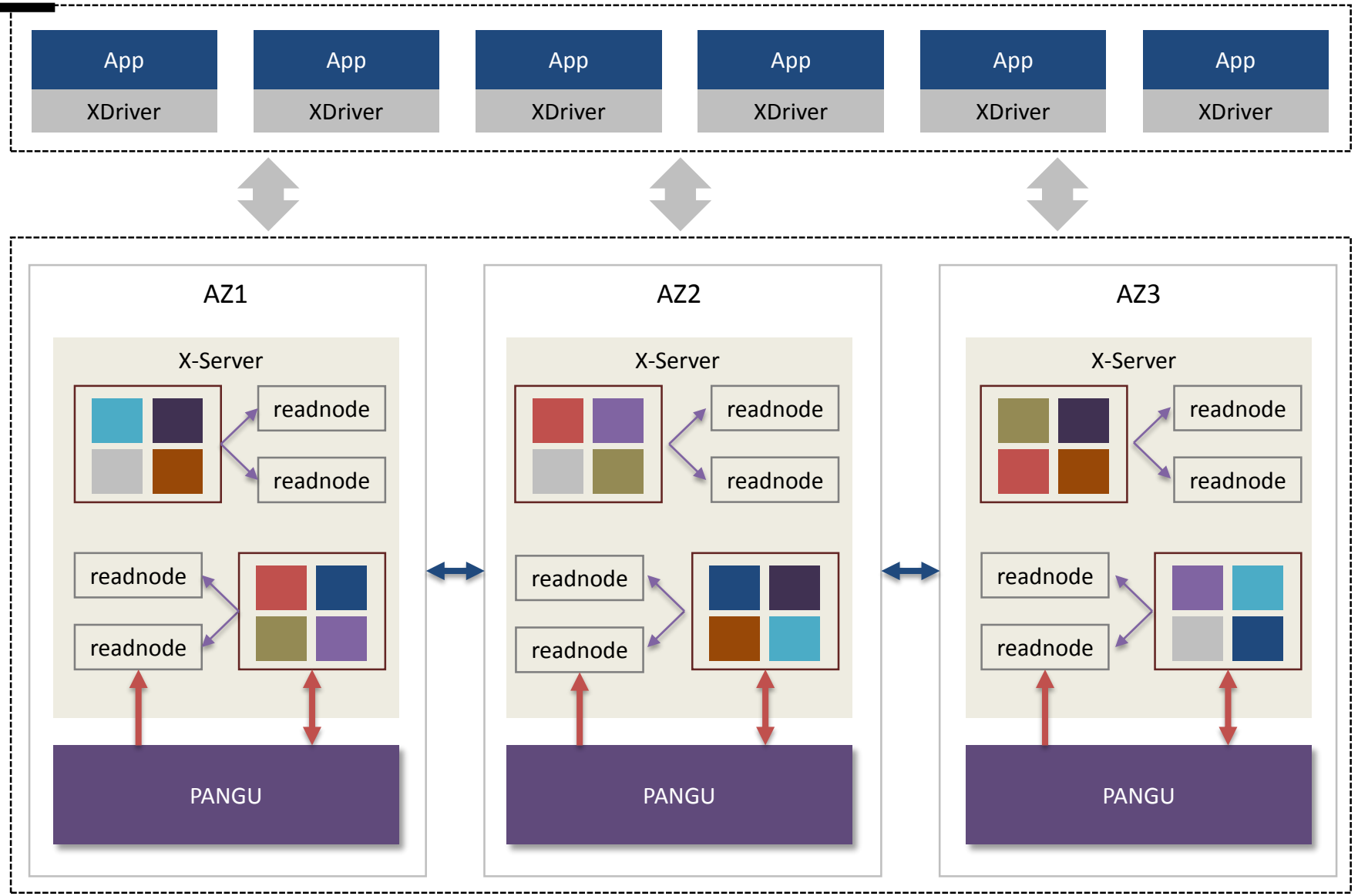
X-DB：架构

X-DB：核心特征

- ✓ 一体化架构，0外部组件依赖
- ✓ 计算、存储水平扩展
- ✓ 数据分片，多点可读可写，降低单点故障影响

X-DB：核心技术

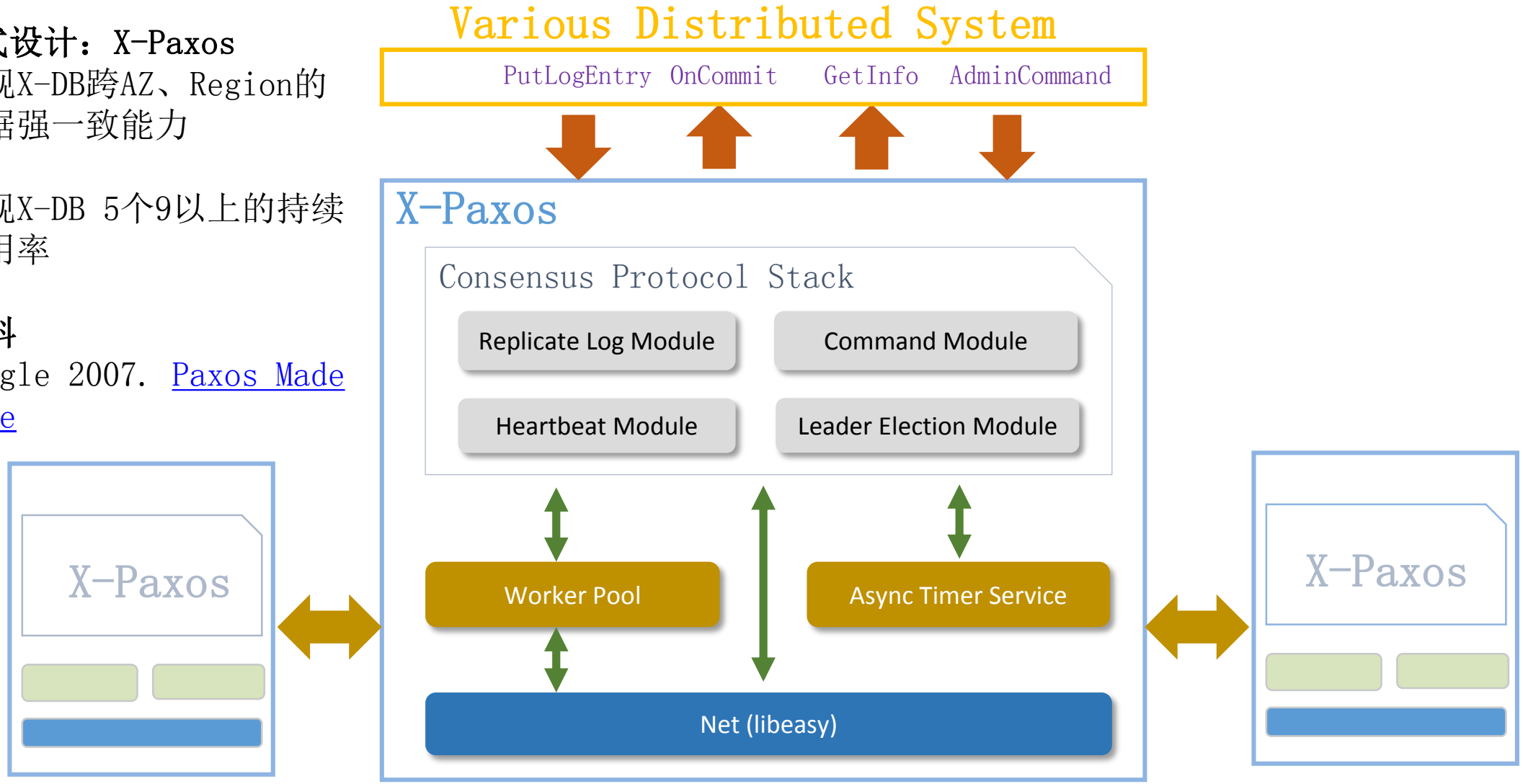
- ✓ X-Paxos
- ✓ X-Engine
- ✓ 计算存储分离
- ✓ SQL Engine



X-DB：核心技术（X-Paxos）

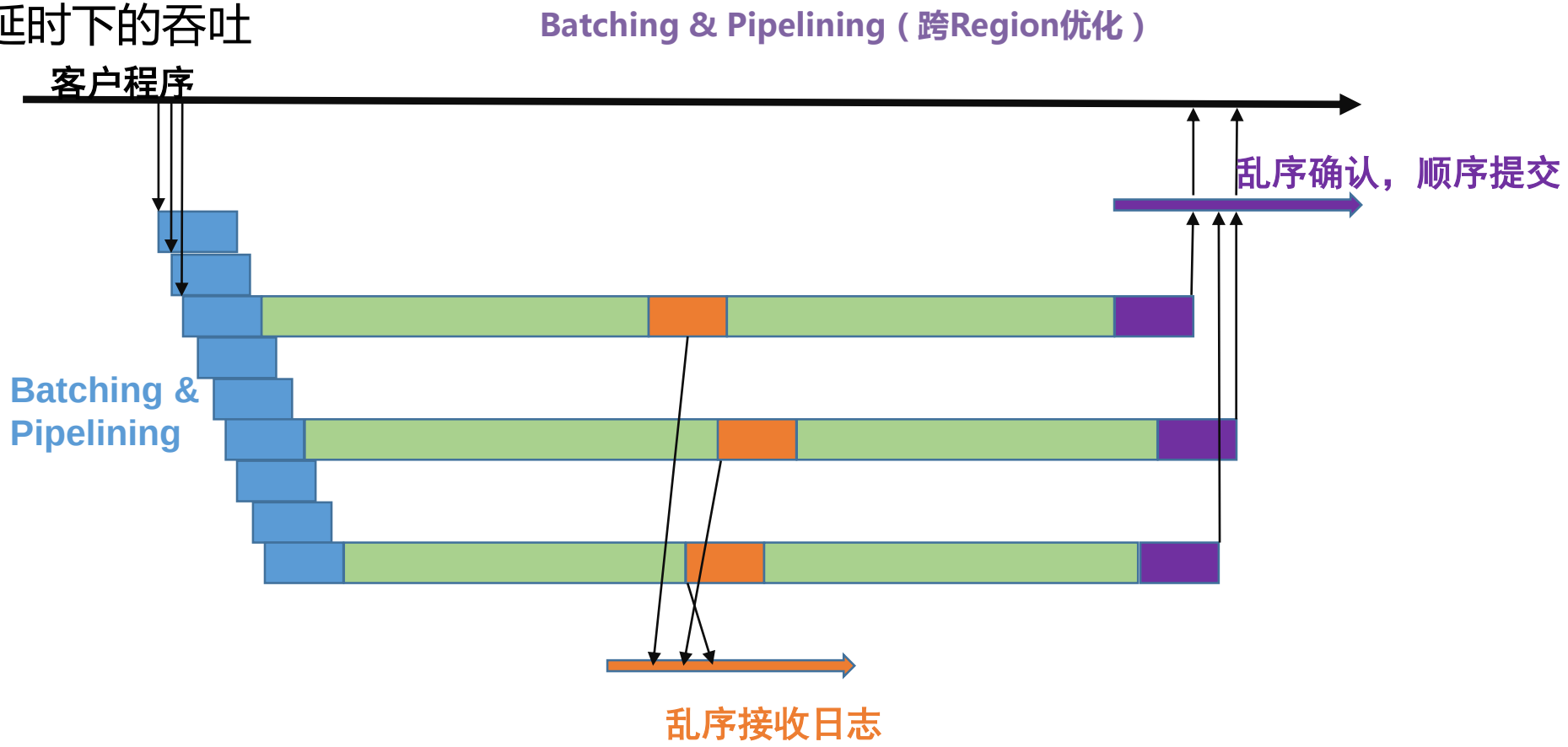
- 为分布式设计：X-Paxos
  - ✓ 实现X-DB跨AZ、Region的数据强一致能力
  - ✓ 实现X-DB 5个9以上的持续可用率

- ✓ 参考资料
  - ✓ Google 2007. Paxos Made Live



X-DB：核心技术（X-Paxos）

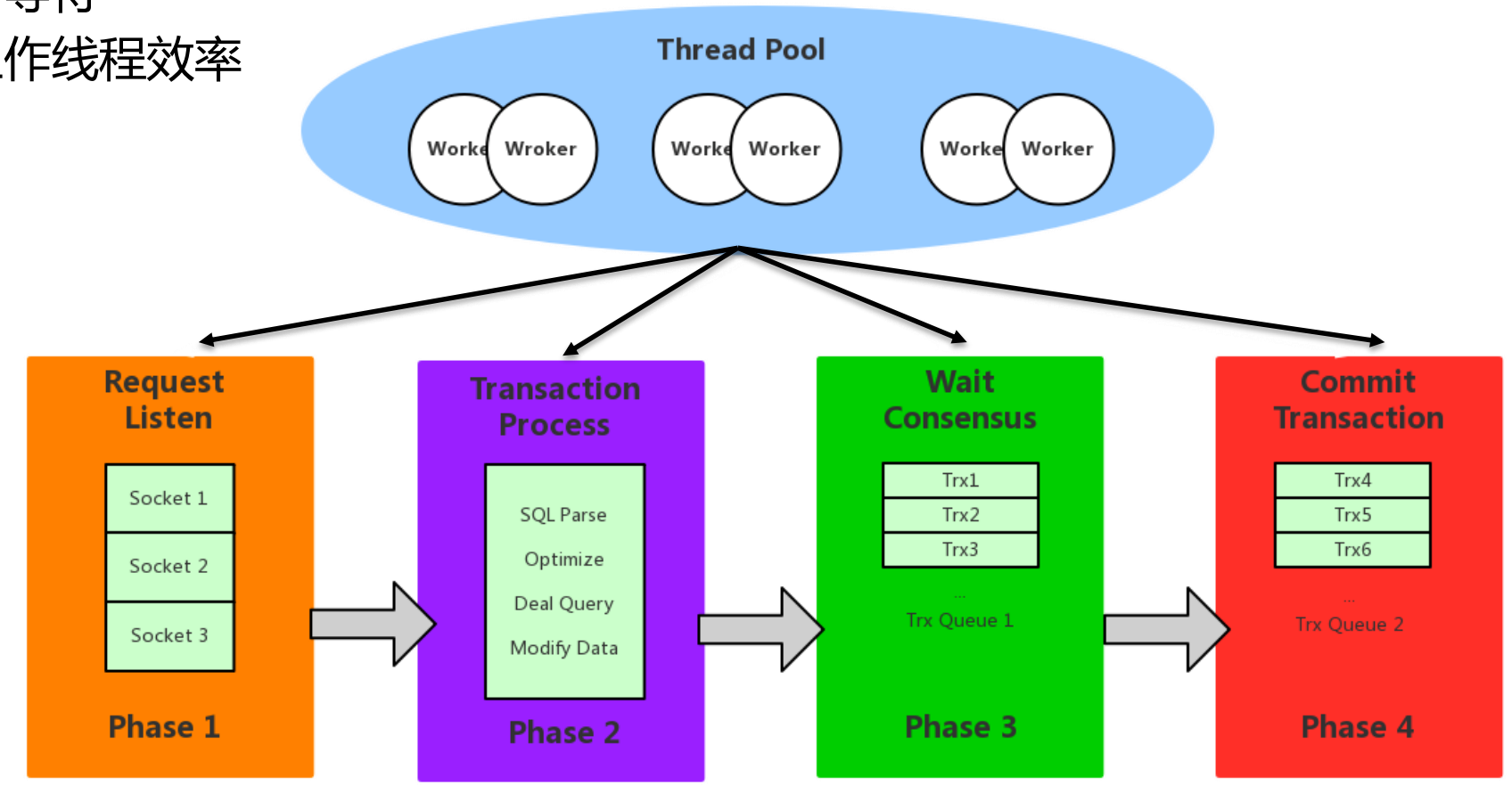
- **Batching & Pipelining**
  - ✓ 提升高延时下的吞吐



- Tuning paxos for high-throughput with batching and pipelining (ICDCN12)

X-DB：核心技术（X-Paxos）

- **异步化**
  - ✓ 消除同步等待
  - ✓ 最大化工作线程效率



X-DB：核心技术（X-Engine）

- **为高性能低成本设计：自研X-Engine存储引擎**
  - ✓ 核心一：数据自动冷热分层
  - ✓ 核心二：基于数据分层架构下的计算和存储优化

- **高性能**
  - ✓ Cache-Conscious Index
  - ✓ Parallel Logging & Recover
  - ✓ NVRAM-Awareness
  - ✓ Lock Free & Latch Free
  - ✓ Adaptive Concurrency Control
  - ✓ …

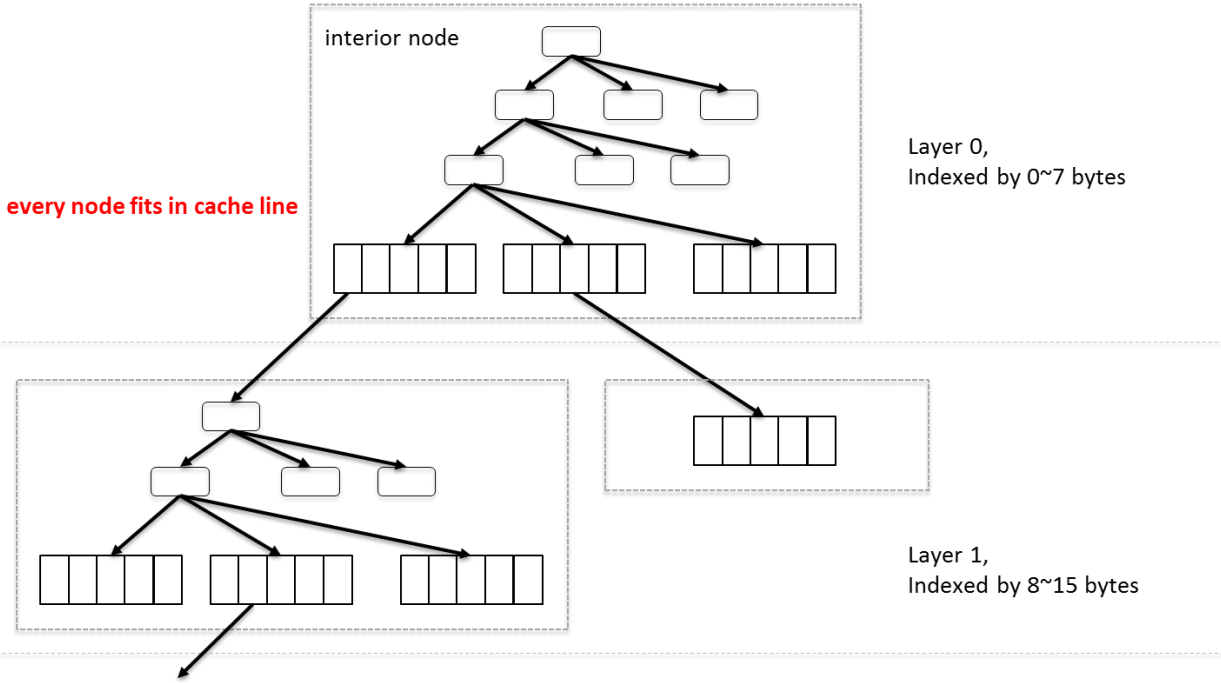
- **低存储成本**
  - ✓ Layered Storage
  - ✓ Adaptive Data Lifecycle Management
  - ✓ Hybrid Row-Column Store
  - ✓ Adaptive Encoding & Compression
  - ✓ …

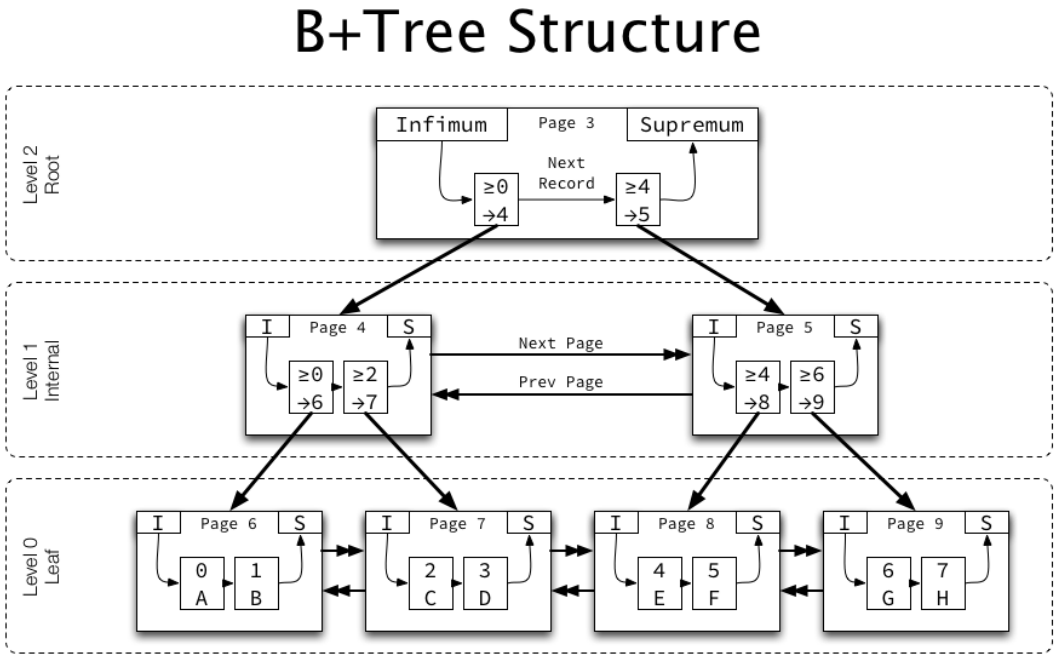
10倍的MySQL写入性能，
百万TPS

MySQL 1/10存储成本

X-DB：核心技术（X-Engine）

高性能内存索引技术





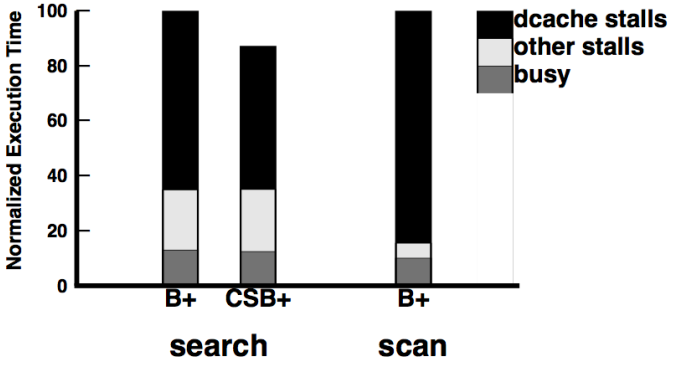
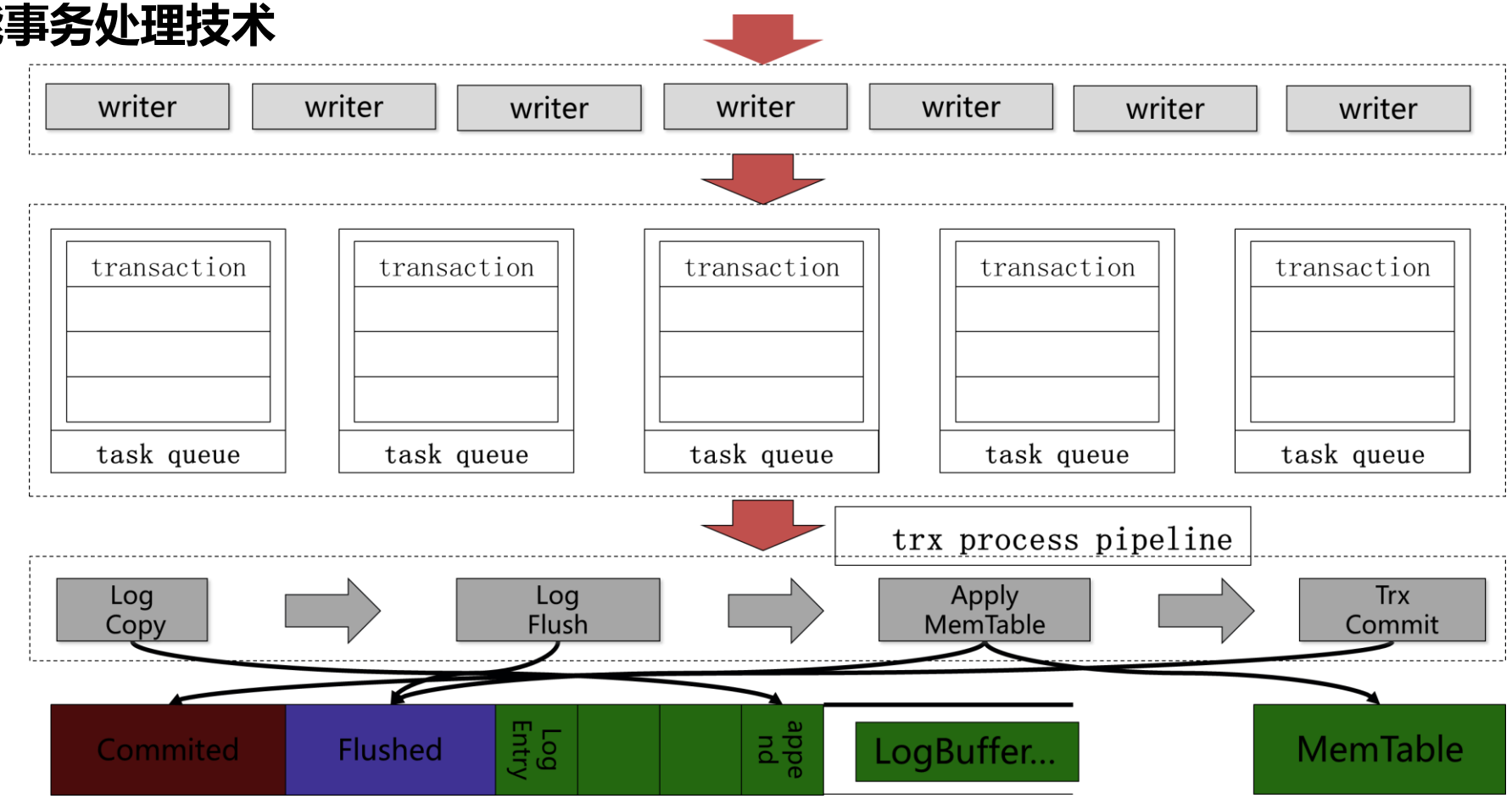


Figure 1: Execution time breakdown for index operations (B+ = B$^+$-Trees, CSB+ = CSB$^+$-Trees).

- 陈世敏. SIGMOD 2001. Improving Index Performance through Prefetching
- MassTree. EuroSys 12. Cache Craftiness for Fast Multicore Key-Value Storage

X-DB：核心技术（X-Engine）

- 高性能事务处理技术

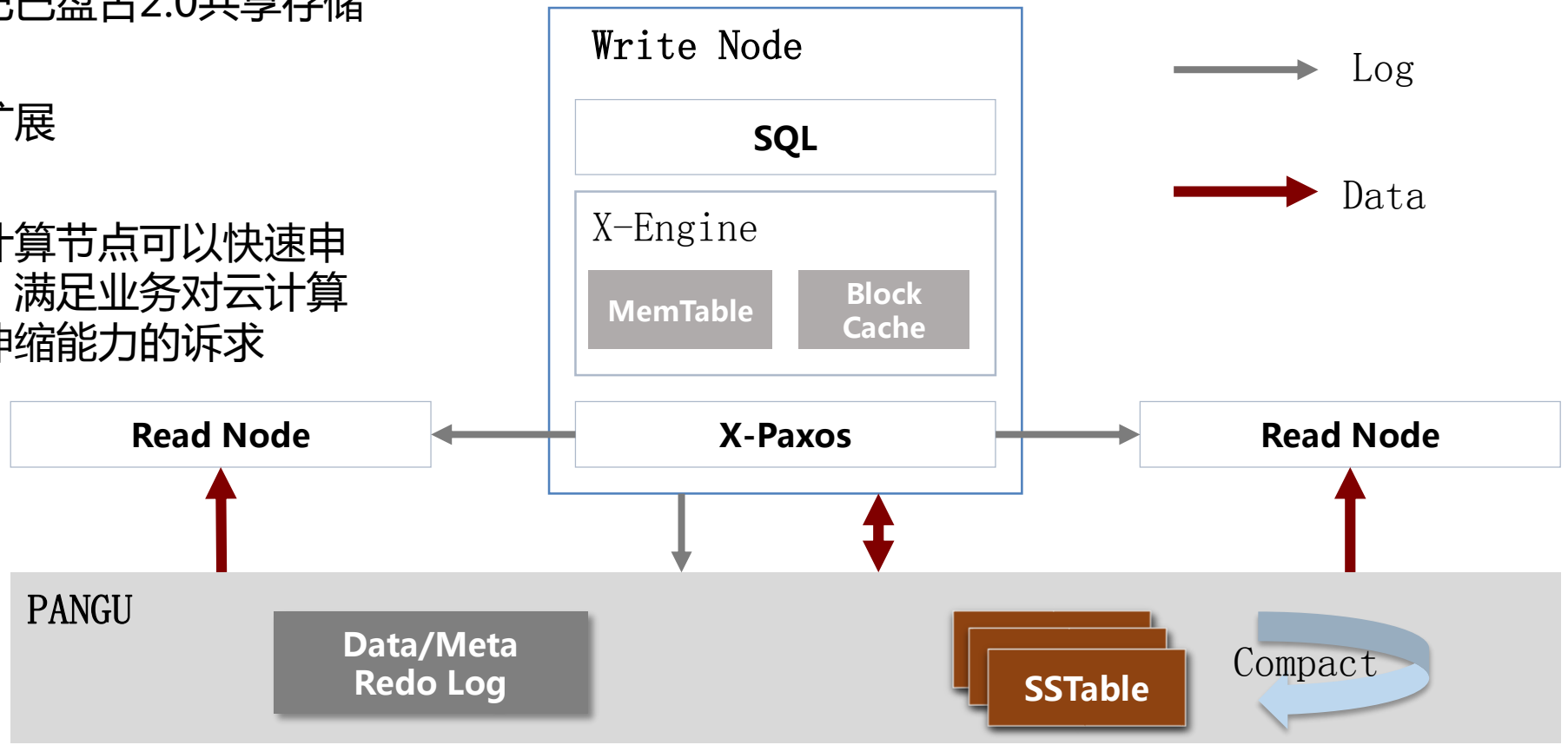


- Ryan Johnson. VLDB 10. [Aether: A Scalable Approach to Logging]

X-DB：核心技术（计算存储分离）

为弹性设计：计算、存储分离

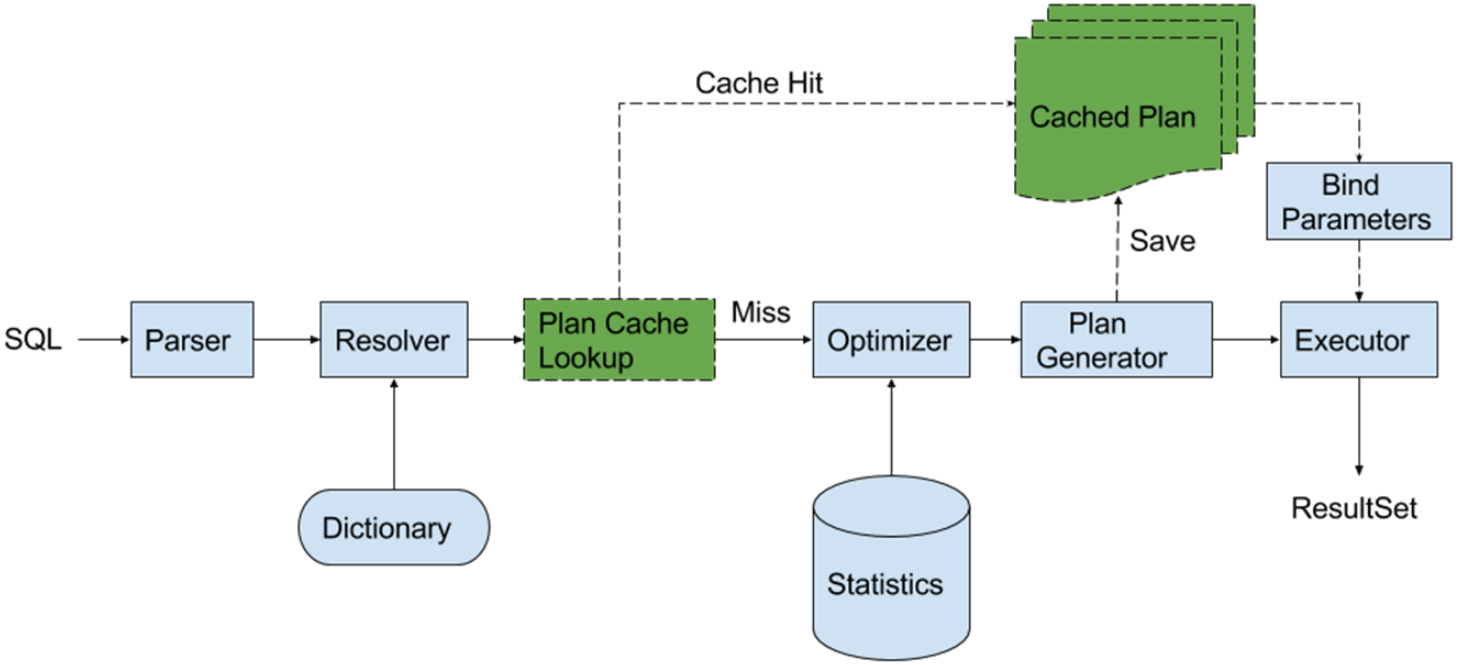
- ✓ 基于阿里巴巴盘古2.0共享存储
- ✓ 存储按需扩展
- ✓ 无状态的计算节点可以快速申请、释放。满足业务对云计算快速弹性伸缩能力的诉求



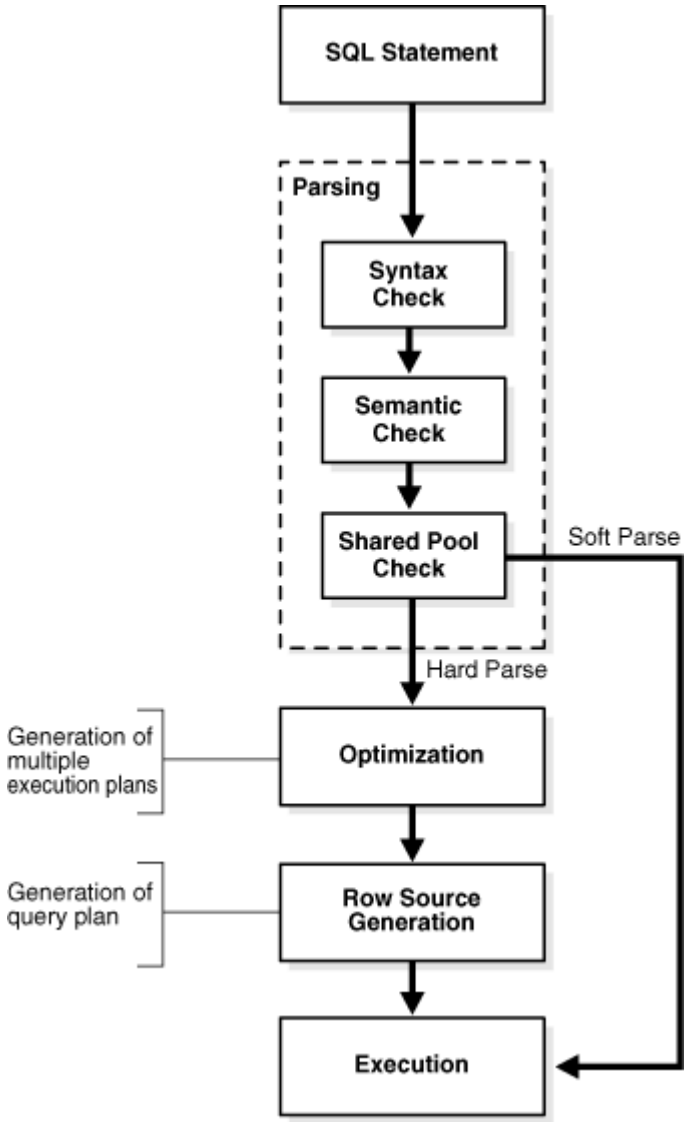
- Amazon Aurora. SIGMOD 2017. Amazon Aurora: Design Considerations for High Throughput Cloud-Native Relational Databases

X-DB：核心技术（ SQL Engine增强）

- **Plan Cache**
  - ✓ 执行计划缓存，弥补MySQL生态最大的不足
  - ✓ 下图：X-DB Plan Cache；右图：Oracle Plan Cache
  - ✓ 性能提升：39%-173%（视不同应用场景）

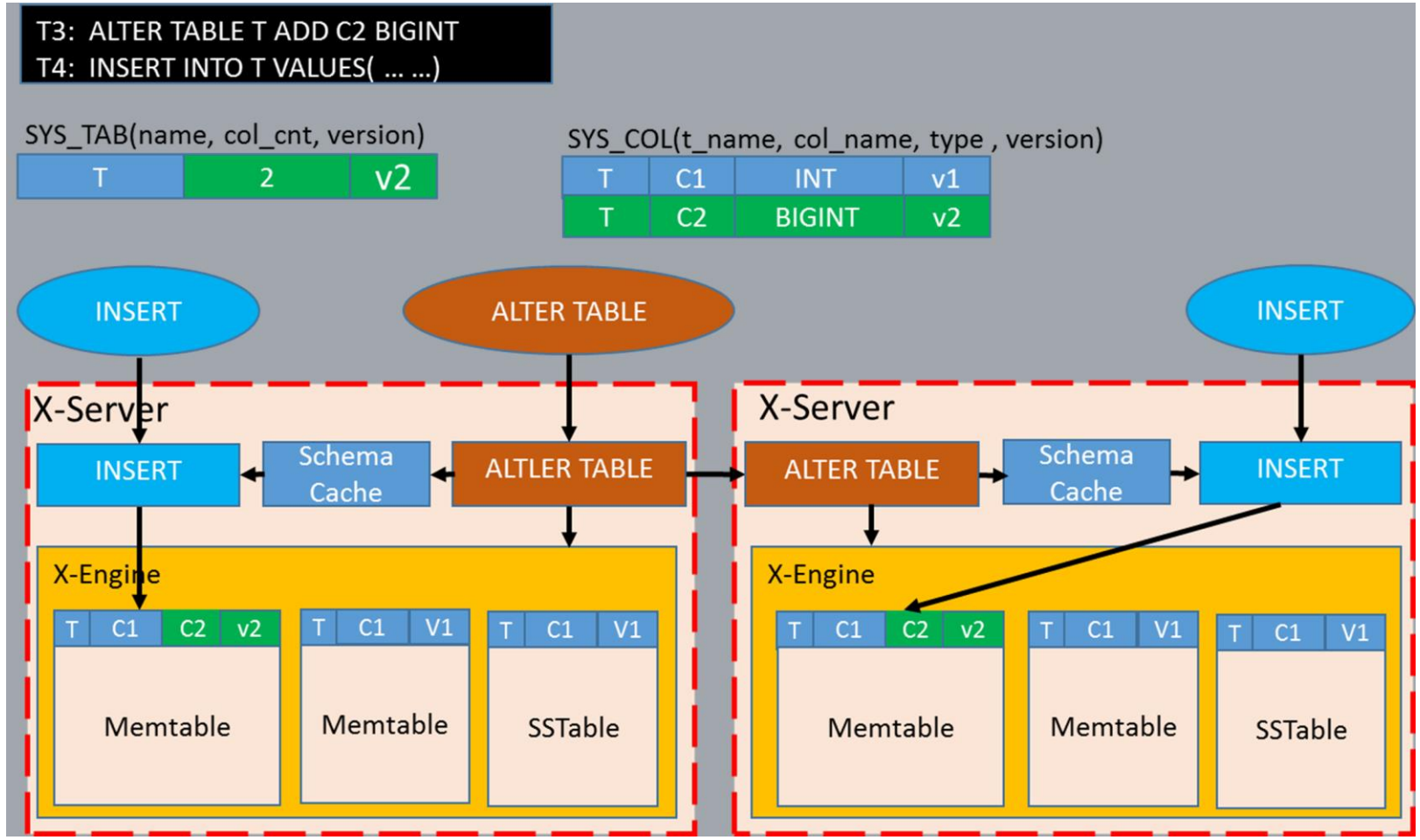


SQL Lifecycle with Plan Cache

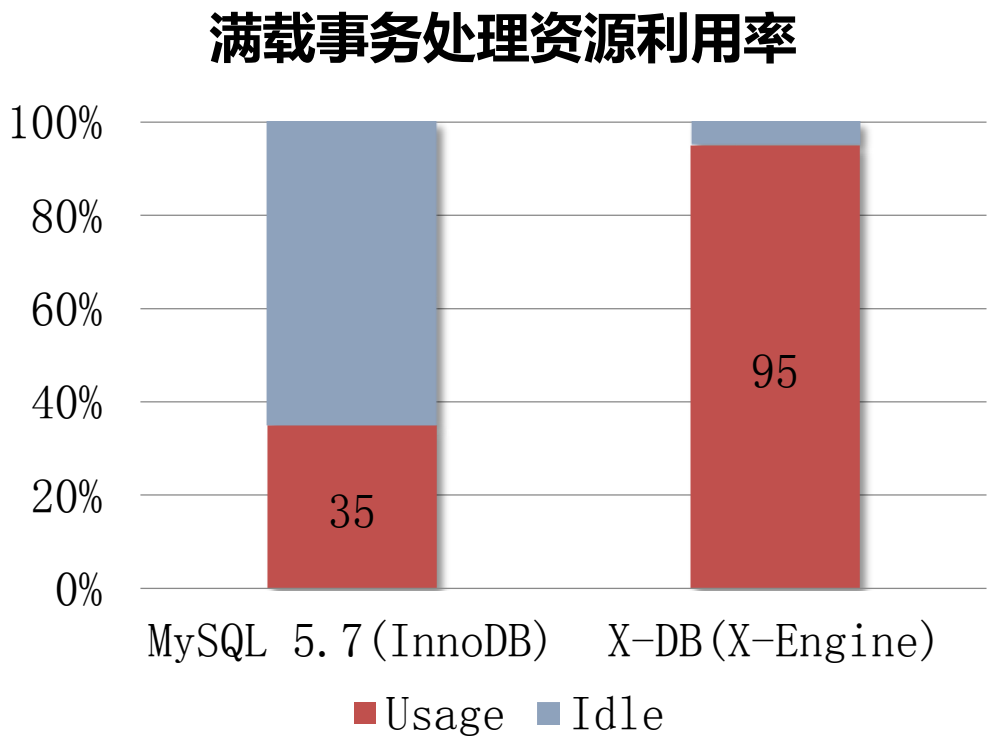


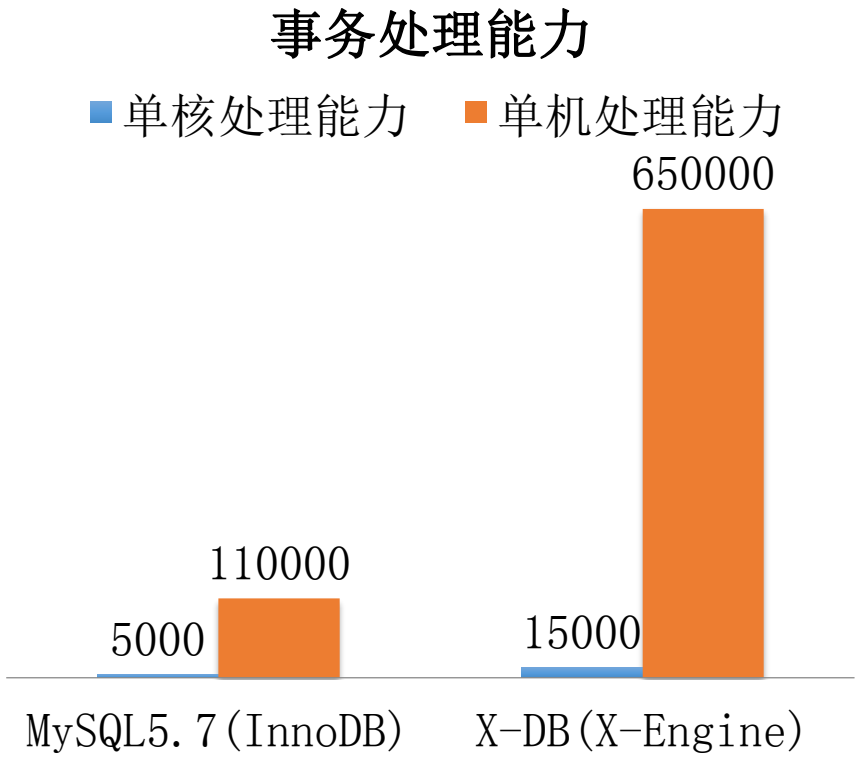
X-DB：核心技术（多版本Schema & Fast DDL）

- **Schema下沉，多版本Schema管理**
  - ✓ 实现真正的Online DDL，No Data Copy
  - ✓ Add/Drop/Modify Column
    - 直接修改元数据，瞬间完成



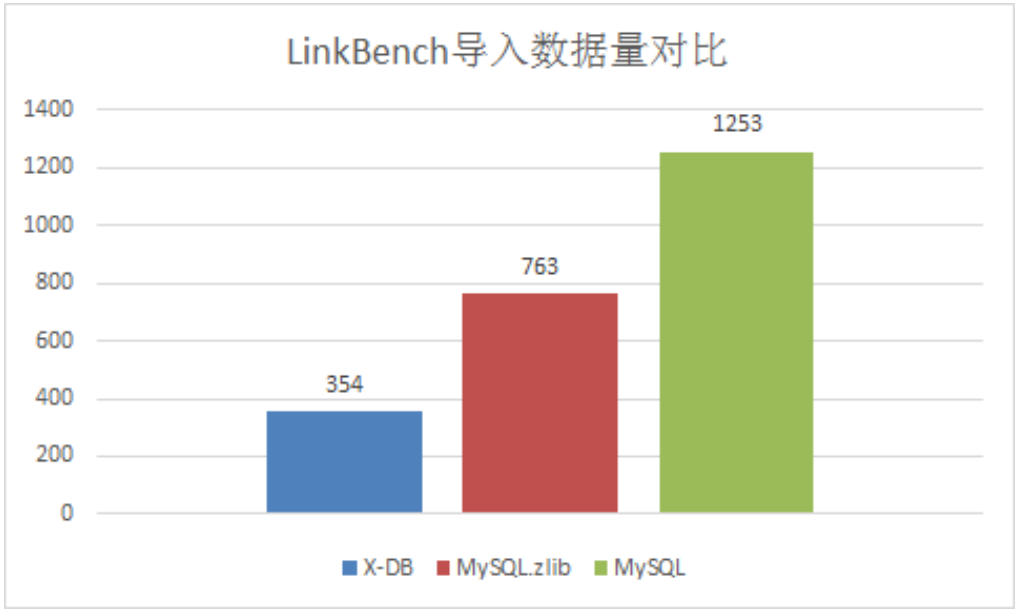
X-DB：Sysbench测试（性能）

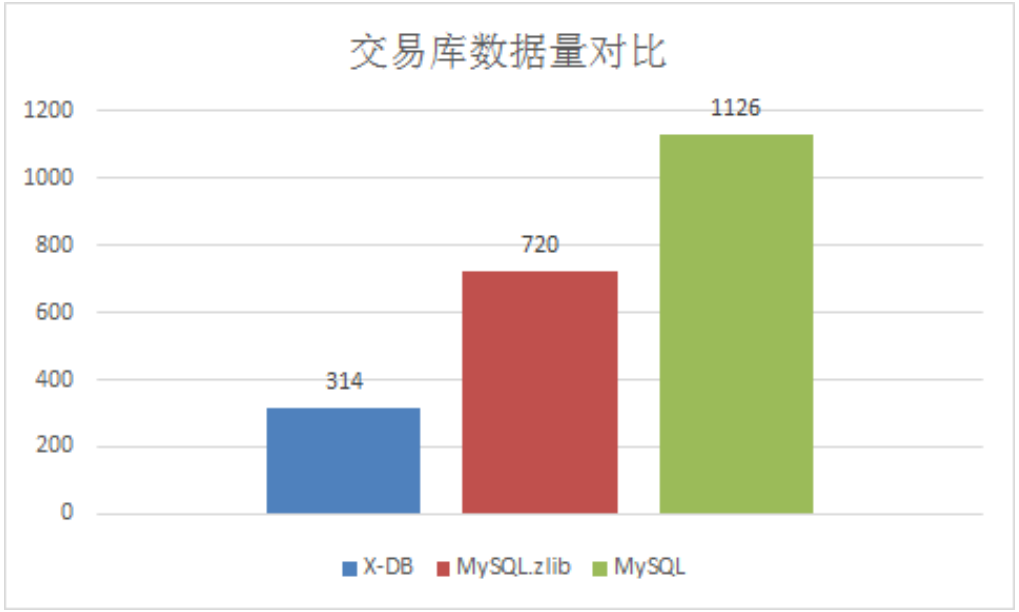




- 真正6倍与MySQL最新版本的性能

X-DB：测试数据（数据压缩）

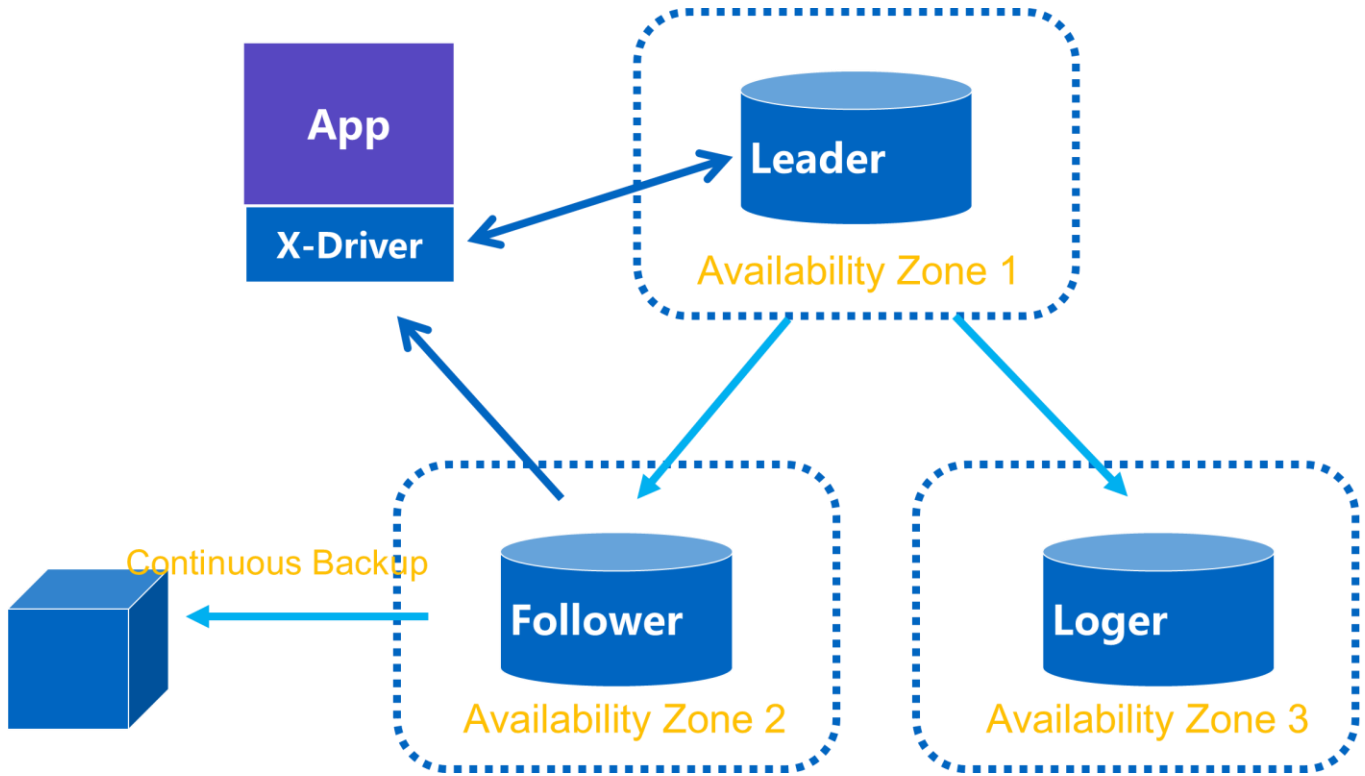




- 目前做到了MySQL 1/4 的存储成本

X-DB：典型应用场景（同城跨AZ部署）

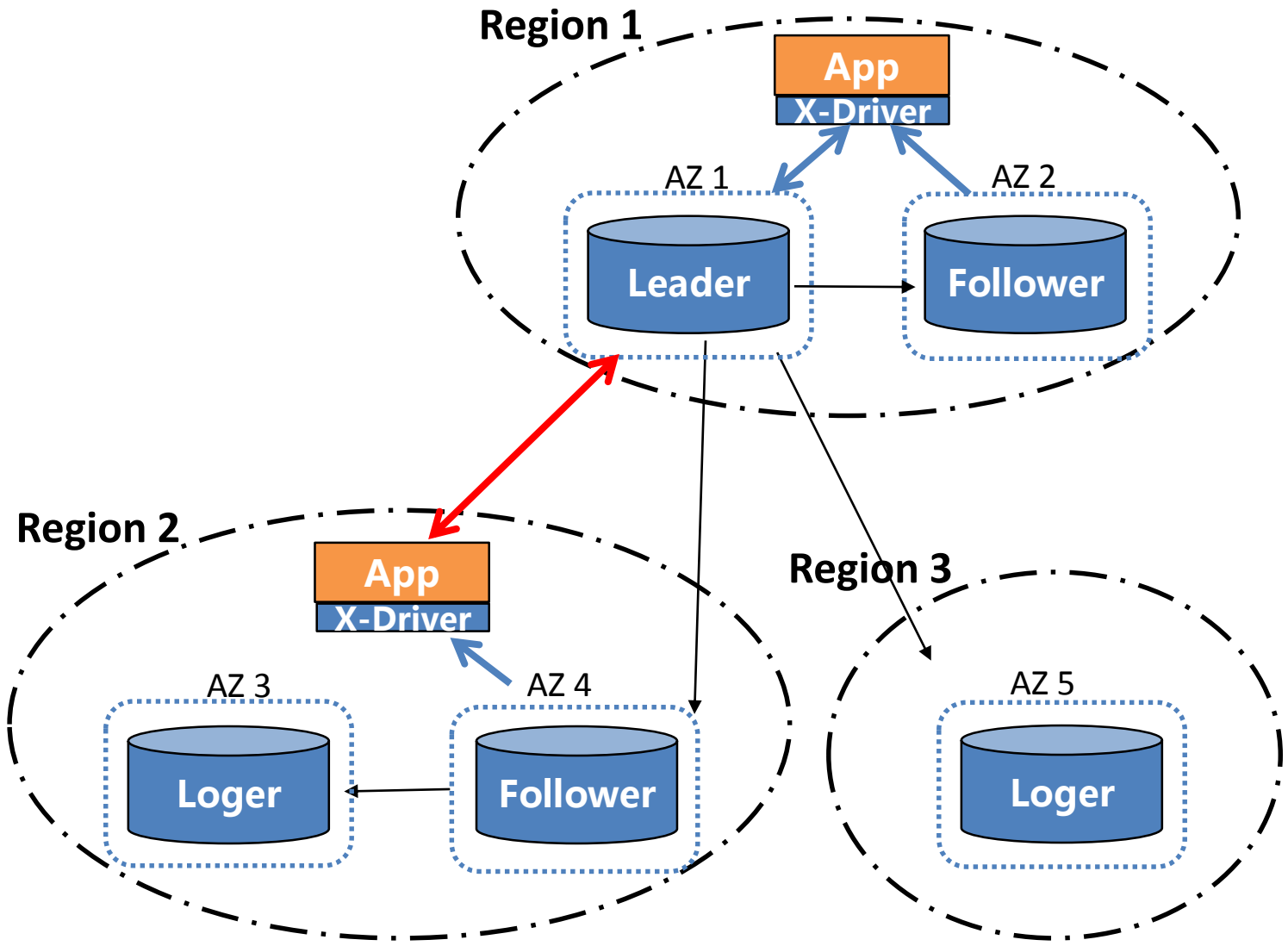
- **数据强一致**
  - ✓ 单AZ不可用数据0丢失
  - ✓ 单AZ不可用秒级切换
  - ✓ 切换自封闭，无第三方组件
- **0成本增加**
  - ✓ 相对主备模式0成本增加
- **持续备份**
  - ✓ RPO < 1s



X-DB：典型应用场景（跨Region部署）

Multi-Region部署

- ✓ 真正Region级的强一致能力
  - 单个Region不可用0数据丢失
- ✓ 高性能
  - 跨Region强同步下依然保持高性能
- ✓ 灵活的切换策略
  - 优先切换同Region
  - 定制跨Region切换顺序
- ✓ 高伸缩性
  - 可无限制的扩充Region/AZ的部署数量和节点数量
  - 可自由的调节Region/AZ内是否部署数据节点，以及数据节点数量



大纲

- 阿里巴巴OLTP数据库应用特点
- 业界数据库发展的关键里程碑
- X-DB：阿里自研高性能分布式数据库
- **阿里巴巴数据库：研究和挑战**

阿里巴巴数据库：研究和挑战

- 他们说，

- Michael Stonebraker
  - ✓ How Hardware Drives The Shape Of Databases To Come

- Wolfgang Lehner
  - ✓ The Data Center under your Desk – How Disruptive is Modern Hardware for DB System Design

- Andy Pavlo
  - ✓ The Next 50 Years of Databases
  - ✓ Building a New Database Management System in Academia

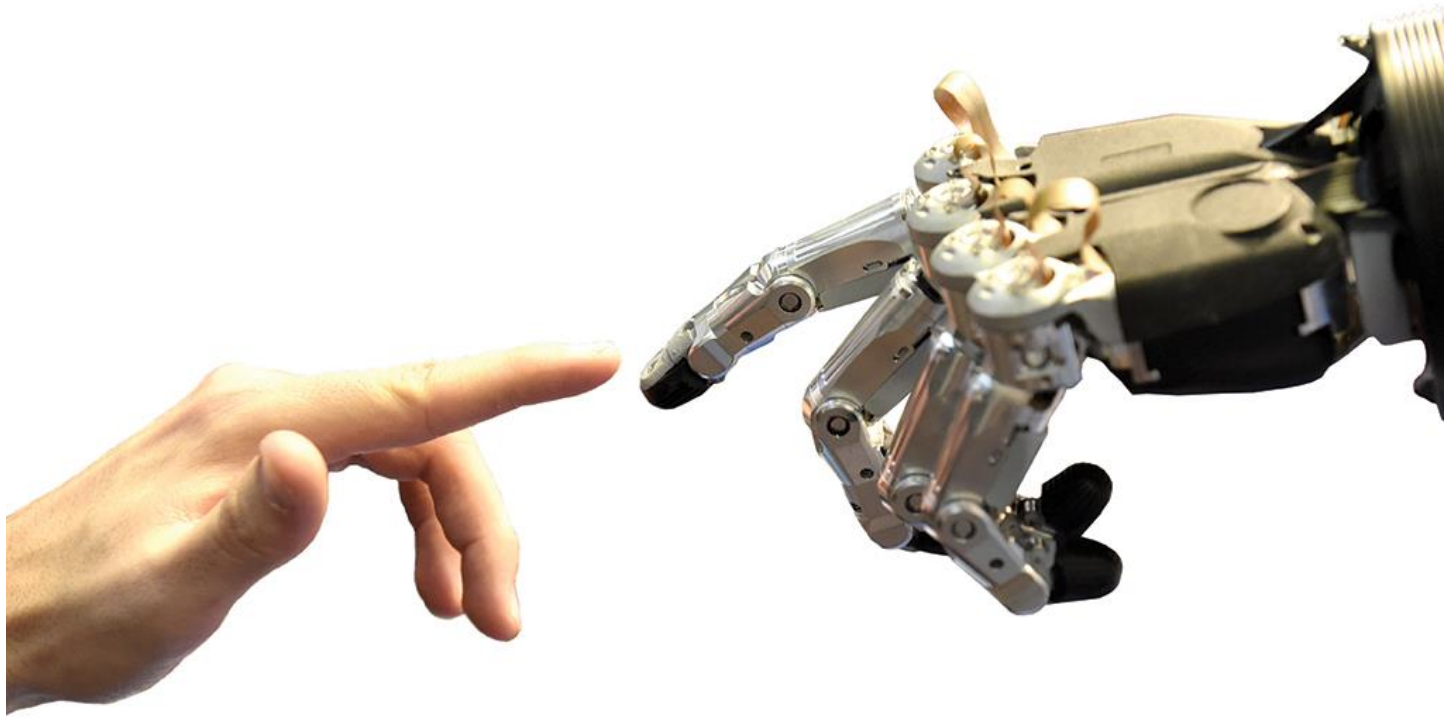
研究与挑战：软硬件Co-Design

- **Multi-Core CPU**
  - ✓ Xiangyao Yu. VLDB 2014. Staring into the Abyss: An Evaluation of Concurrency Control with One Thousand Cores
- **Heterogeneous Computing**
  - ✓ Kaan kara. 2017. FPGA-based Data Partitioning
- **NVRAM**
  - ✓ Andy Pavlo. 2016. What Non-Volatile Memory Means for the Future of Database Systems
  - ✓ Hideaki Kimura. 2015. FOEDUS: OLTP Engine for a Thousand Cores and NVRAM
- **High Performance Network**
  - ✓ Aleksandar Dragojevic. NSDI 2014. FaRM: Fast Remote Memory

研究与挑战：Hybrid transactional/analytical processing (HTAP)

- What is HTAP ?
  - ✓ Hybrid Transactional/Analytical Processing
  - ✓ *Hybrid transaction/analytical processing (HTAP) is an emerging application architecture that "breaks the wall" between transaction processing and analytics. It enables more informed and "in business real time" decision making.*
- 代表产品
  - ✓ TUM Hyper： A Hybrid OLTP&OLAP High Performance DBMS
  - ✓ SAP HANA
  - ✓ CMU Peloton

研究与挑战：数据库与机器学习



- Dana Van Aken. AWS AI Blog 2017. Tuning Your DBMS Automatically with Machine Learning
- Dana Van Aken. SIGMOD 2017. Automatic Database Management System Tuning Through Large-scale Machine Learning

研究与挑战：重定义数据库标准化测试

- 业界数据库测试标准
  - ✓ TPCC：1992年7月发布
- 近年来新的测试工具
  - ✓ Yahoo!. YCSB. SoCC 10. Benchmarking Cloud Serving Systems with YCSB
  - ✓ Facebook. Linkbench. SIGMOD 13. LinkBench: a Database Benchmark Based on the Facebook Social Graph
- 思考ImageNet
  - ✓ Li Fei-Fei. CVPR09. ImageNet: A Large-Scale Hierarchical Image Database
  - ✓ 基于阿里世界最大的在线交易、支付系统平台，Re-Define业界标准！
  - ✓ 我们走出的第一步：史无前例开放！阿里内部集群管理系统Sigma混布数据

写在最后

- 数据库：三大基础软件系统之一。在新的时代，面临着新的契机与挑战！
- 希望有更多的同学能够参与到这个挑战中来。阿里巴巴达摩院
- 个人比较喜欢的几本数据库书籍
  - ✓ Stonebraker. Architecture of a Database System
  - ✓ Stonebraker. Readings in Database Systems, 5th Edition
  - ✓ Jonathan Lewis. Oracle Core: Essential Internals for DBAs and Developers
- 关心新型数据库的几篇综述Paper
  - ✓ Andrew Pavlo. What's Really New with NewSQL
  - ✓ Hao Zhang. In-Memory Big Data Management and Processing: A Survey

谢谢！