DTCC

第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

架构革新 高效可控

IT168.com ChinaUnix.net ITPUB

北京国际会议中心 | 2020/12/21-12/23

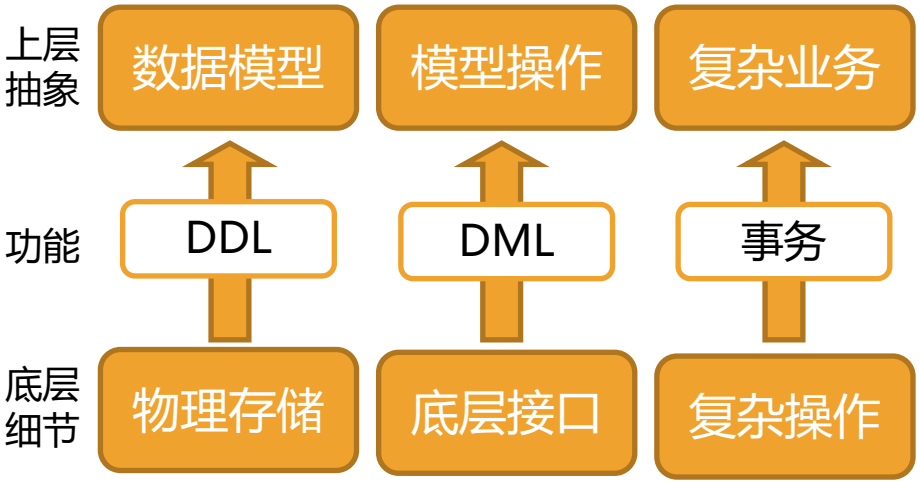
OCEANBASE

深入 OceanBase
企业级数据库的分布式事务引擎

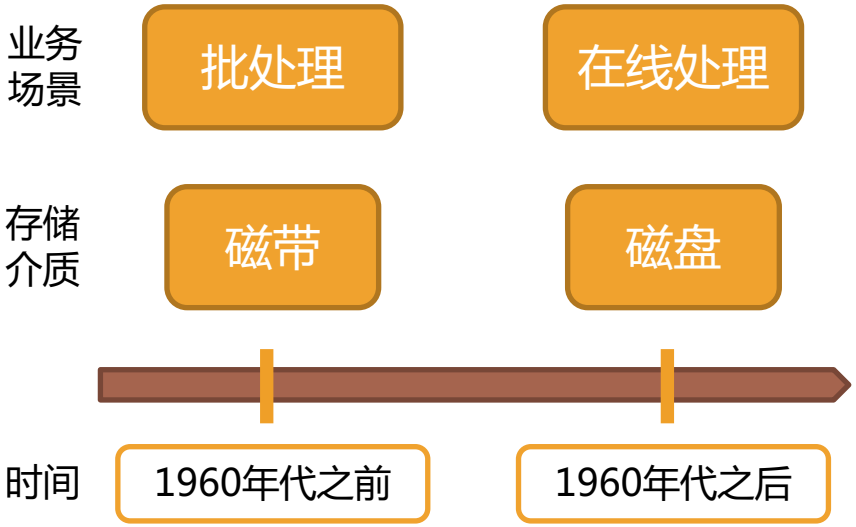
韩富晟（颜然）

北京奥星贝斯科技有限公司资深技术专家

什么是事务？



- 数据库管理系统的最大价值是通过抽象简化业务构建
- 用**数据模型**抽象表达硬盘上存储的裸数据
  - 层次、网络、关系、对象
- 用**模型上的操作**抽象表达存储设备的 IO 操作
  - Create Read Update Delete
- 用**事务模型**抽象表达业务逻辑



- 1965年 Sabre系统上线
- 数据库管理系统作为更多“关键业务”的底座应用在方方面面
- 数据库管理系统作为一个独立的系统在60年代末被广泛接受

事务处理系统与数据库管理系统

事务型中间件

数据库管理系统

操作系统

事务处理系统
Transaction Processing System

- 事务处理系统面向业务，解决实际业务的问题
- 保证最终的业务操作具有事务特性

事务处理系统与数据库管理系统

- 数据库管理系统的面向抽象后的数据模型，抽象程度更高
- 事务特性面向数据模型上的操作

数据库管理系统
DataBase Management System

事务抽象的是什么？

Atomicity　Durability　Isolation　Consistency

故障恢复 Failure Recovery

- 持久性（Durability）为什么难？
- 为计算机编制好面对一切情况的手册

并发控制 Concurrency Control

- 从批处理到在线服务的最大变化
- 计算机资源利用率最大化是复杂性的来源

故障恢复

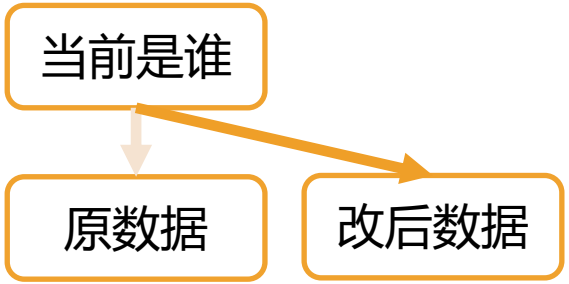
- 日常生活中故障也经常发生
  - 考试时写字的笔没水了
  - 超市付款时手机没电了
  - ……
- 人可以轻松应对这些问题

- 计算机需要编制好程序告诉它该怎么做。需要为它编制好面对一切情况的手册

- 实际系统中有两个得以广泛运用的方法：
  - Redo Logging
  - Shadow Paging

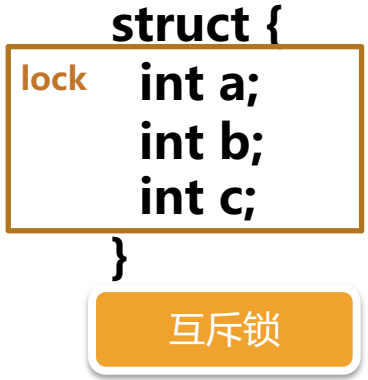
Redo Logging:

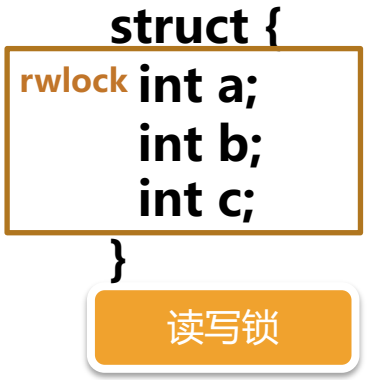
Shadow Paging:

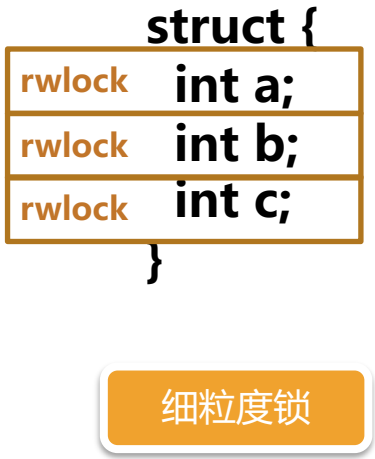


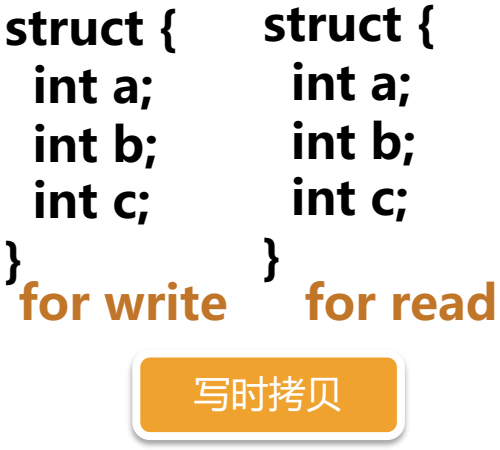
并发控制

- 所有程序都会面对的挑战
- 并发控制的目的是保证正确性的同时让计算机资源利用率最大化
- 资源利用效率由低到高：
  - 互斥锁
  - 读写锁
  - 细粒度锁
  - 写时拷贝



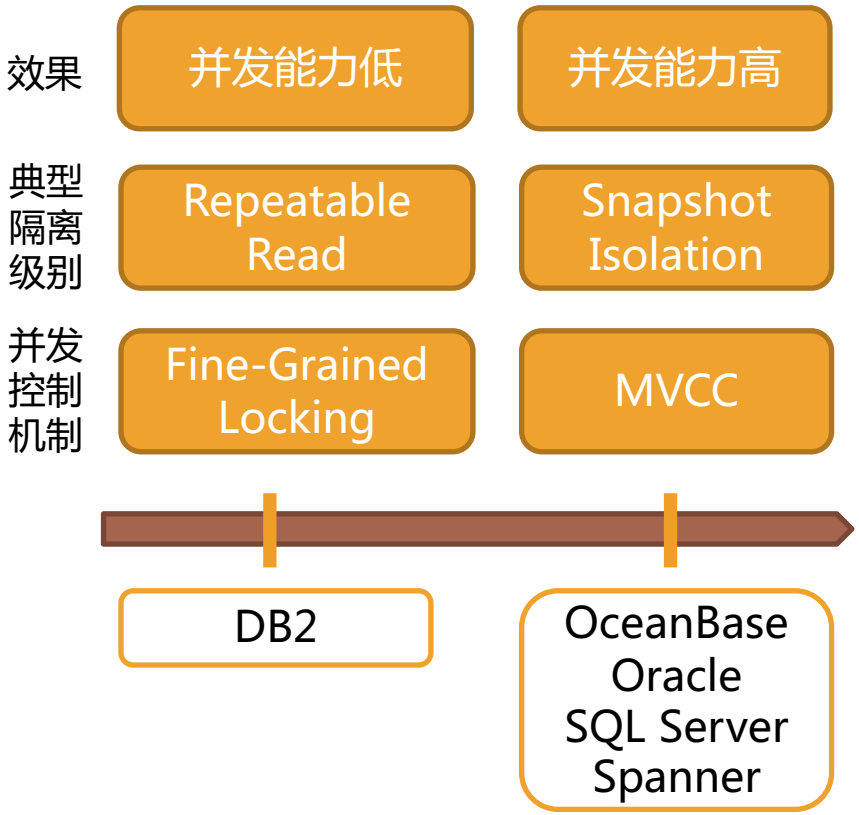






并发控制

- 并发控制机制：从基于锁的机制到多版本并发控制机制
- 多版本并发控制机制
  - 最好的并发能力
  - 读取操作不影响写入
  - 更适合并发越来越大的分布式数据库
- 大量的数据库系统都采用多版本并发控制机制



并发控制

Snapshot : Read View

No Commit Version

Snapshot : Read Version

Have Commit Version

Read View:

ID<=90

ID=92

ID=94

ID=95

! ID>95

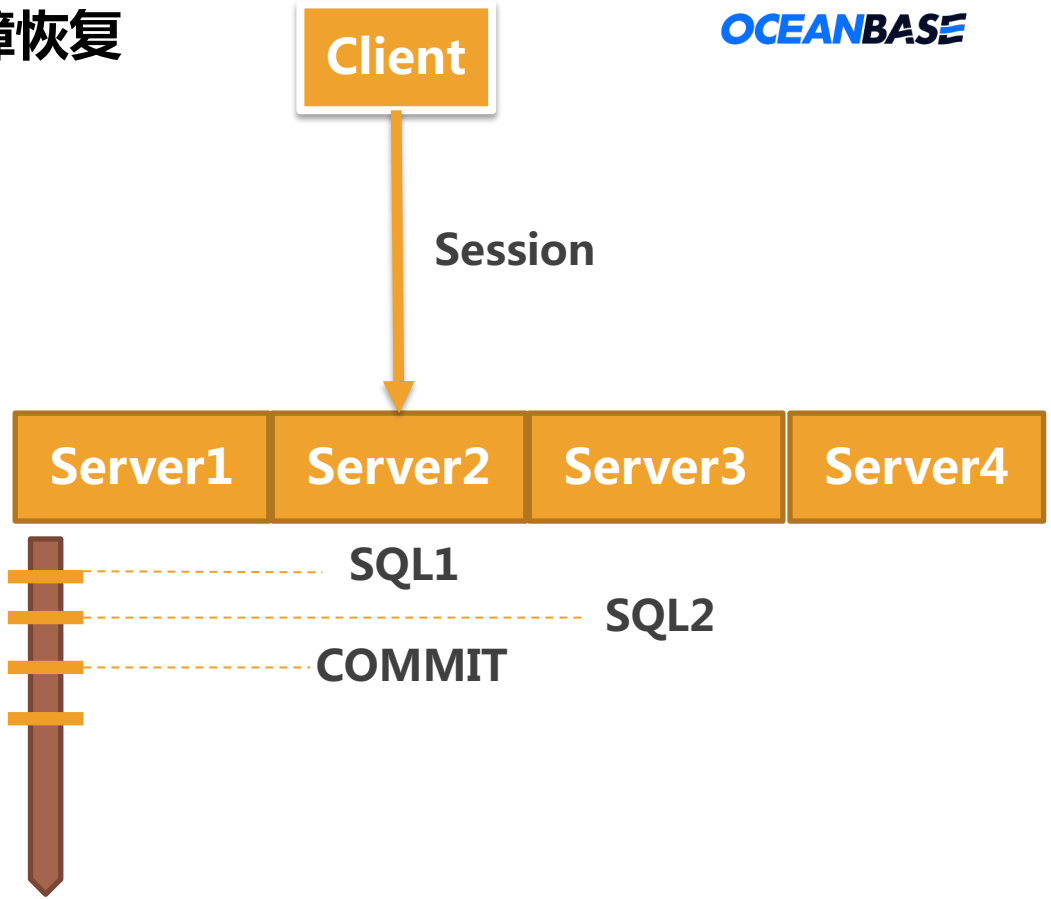
- Read View：记录快照点所有活跃事务列表
  - 取快照点非常耗时，不具备扩展性
- Read Version：快照版本号
  - 快照点是一个值，支持高并发获取

Read Version:

Version=200

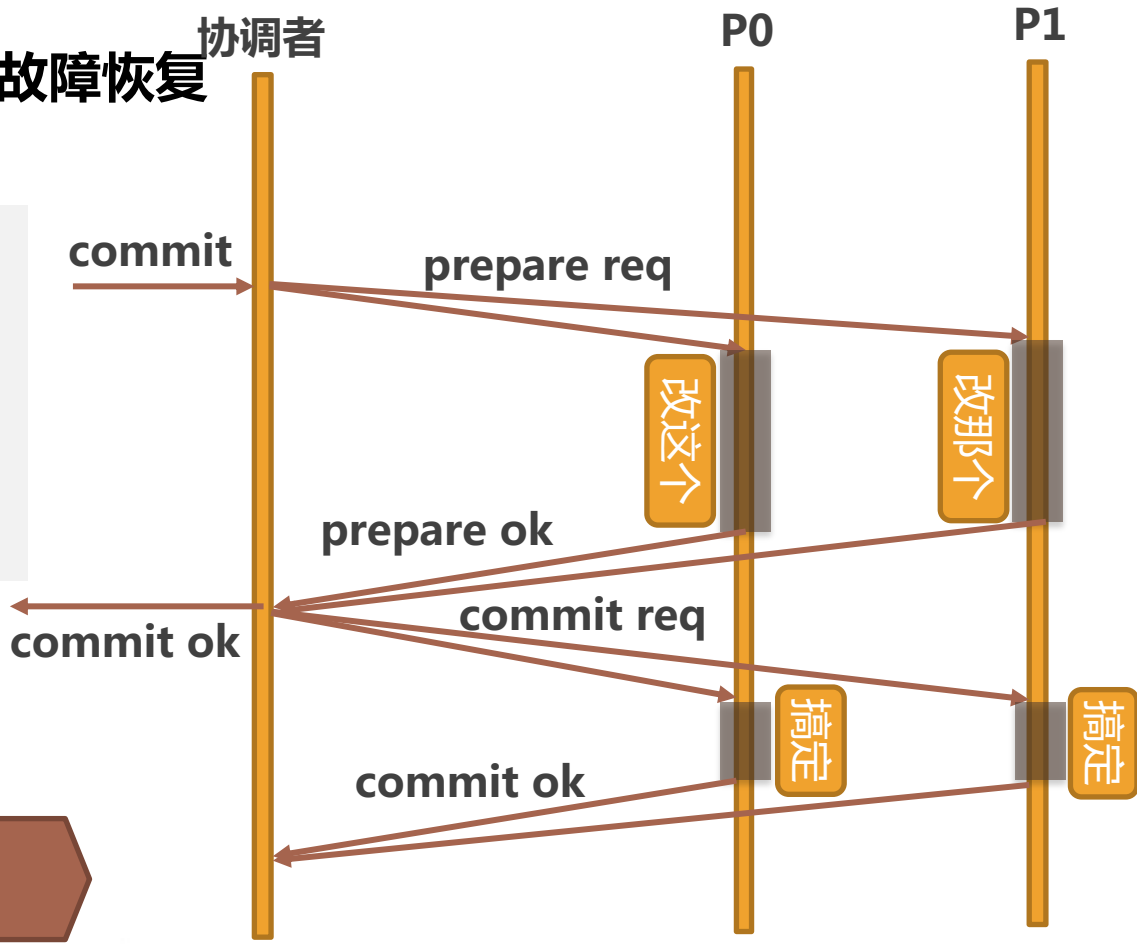
分布式环境下的故障恢复

- 分布式环境下的故障恢复
  - 同时修改多台机器的数据，原子性的保证更挑战
- 自动的分布式提交
  - 系统内部自动记录数据修改发生的位置
  - 事务提交时自动选择走一阶段提交还是两阶段提交
  - 保证跨机事务的原子性
  - 低延迟



分布式环境下的故障恢复

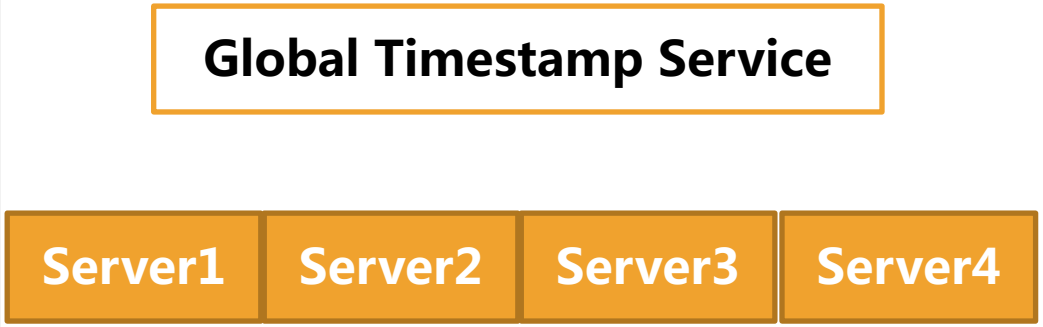
- OceanBase 分布式事务协调者无持久化状态
- OceanBase 两阶段提交低延迟：应用Commit请求只有1次日志延迟



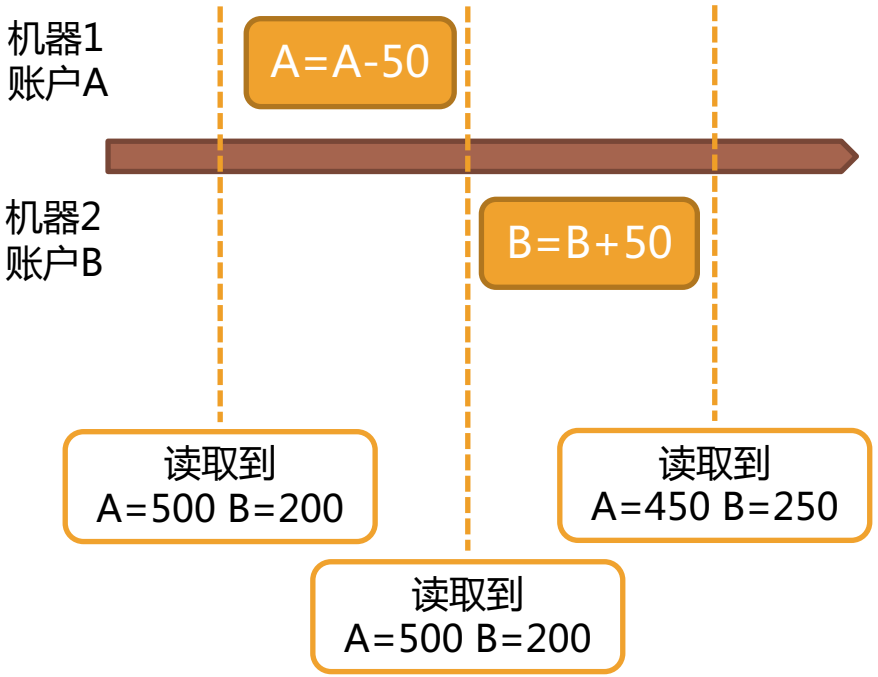
Redo Logging:

分布式环境下的并发控制

- 分布式环境下的并发控制
  - 要协调多台机器的不同操作
  - 跨越多台机器的读取要有一致性的快照
- 高可用GTS服务
  - 多副本高可用
  - 性能强
  - 高效的聚合能力
  - 支撑TPC-C测试每分钟15亿事务



分布式环境下的并发控制



- 高效的未提交数据存储能力
  - 事务执行过程中产生的修改都以未提交数据存储在系统中
  - 读取根据快照版本选择系统中的历史数据
    - 不会见到正在修改的数据
    - 只会见到版本一致的数据

OceanBase事务特性

- INSERT INTO RES
  SELECT C1, C2, C3, C4
  FROM A, B
  WHERE A.NO = B.NO
  AND B.TYPE != 'XX'

- UPDATE A
  SET STATUS = 1
  WHERE A.EV_DATE
  BETWEEN
  TO_DATE('2008-JUN-01', 'YYYY-MON-DD')
  AND
  TO_DATE('2008-JUL-01', 'YYYY-MON-DD')

不限制事务大小

修改与读取隔离

批量写入保证原子性

提交瞬间生效

OceanBase事务特性

- A C I D
- Savepoint/Nested Transaction
- XA
- ...

总结

- OceanBase 是原生分布式数据库，利用分布式集群给用户提供可扩展和高可用的数据库系统服务
- OceanBase 给用户透明的体验，像使用单机数据库一样使用分布式数据库，没有分库分表的烦恼，支持完整的事务功能

OceanBase 微信公众号

OceanBase 官网

THANKS