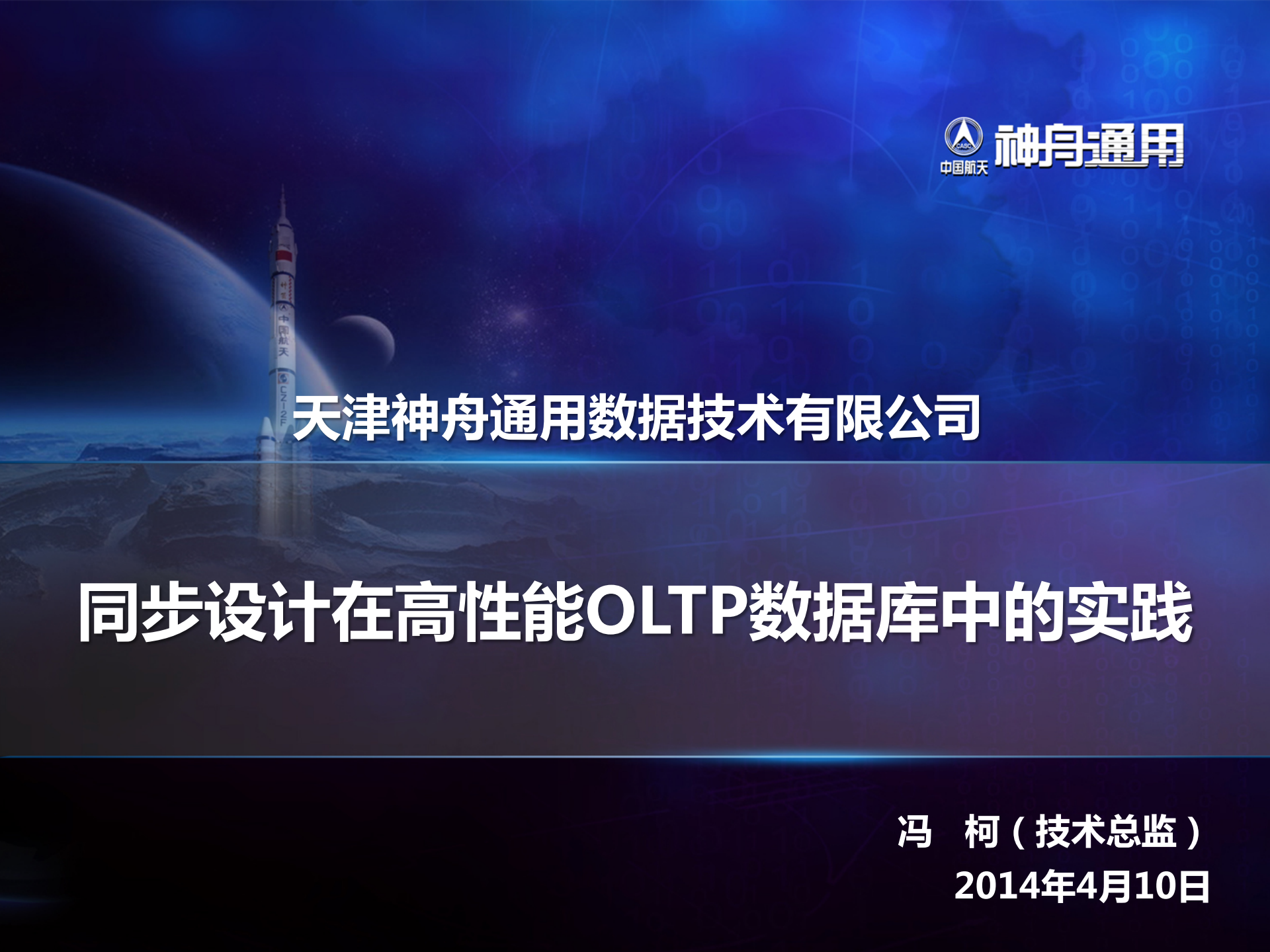

中国航天
神舟通用
天津神舟通用数据技术有限公司
同步设计在高性能OLTP数据库中的实践
冯　柯（技术总监）
2014年4月10日

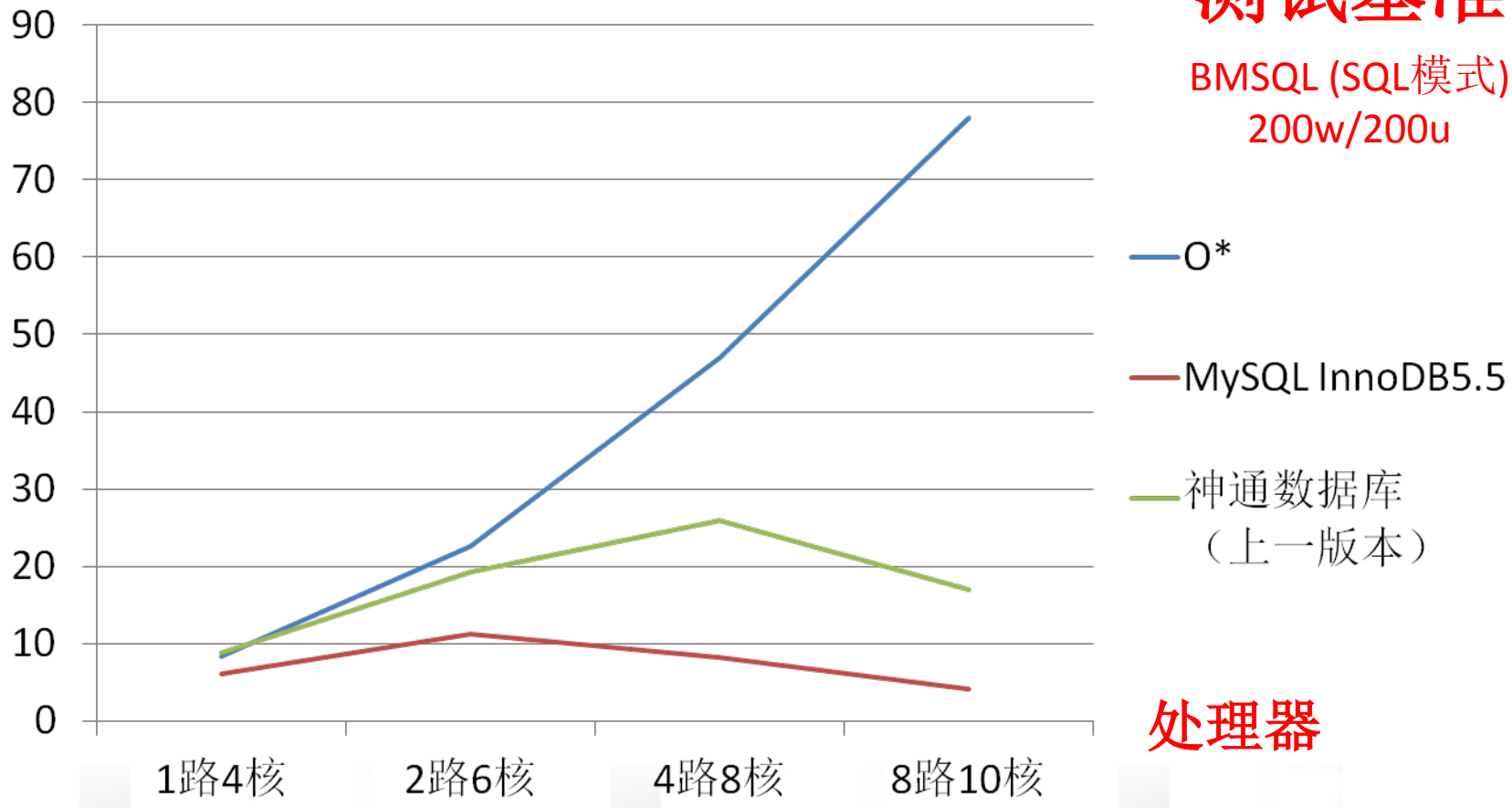
关于OLTP

- **数据库界的珠穆朗玛**
- **贫富差距严重**
- **应用特征**
  - 高并发
  - 短事务
  - 更新密集

OLTP vs. OLAP

业界现状

单位：万笔交易/分钟



TPC-C
测试基准

BMSQL (SQL模式)
200w/200u

处理器

问题出在哪？

- **并发协议出了问题吗？**
- **我们的分析：**
  - 并发协议是整个并行系统的基础
  - 优秀的并发协议（如MVCC读写并行）对高性能OLTP数据库至关重要
  - 但是：好的协议 ≠ 好的实现（测试使用的三个数据库均支持MVCC！）

问题出在哪？（2）

- **同步原语出了问题吗？**
- **我们的分析：**
  - 同步原语实现的目标，在于减少同步本身的开销，在临界区执行路径较短时，可以发挥很大作用
  - 以数据库为代表的复杂并行系统，真正决定并发性能的，往往是一些逻辑复杂、执行路径较长的临界区，优化同步本身对性能的帮助有限

同步原语一览

类Latch的实现(Ticket/Queue…)

Atomic 原子操作(Counting…)

Lock-Free结构 (FIFO/FILO/HASH…)

Memory Barriers (Sequence Locking…)

RCU…

…

问题出在哪？（3）

- 主要现象：关键临界区争用剧烈，部分临界区出现“越并发性能越差”问题
- 解决问题的关键在于优化同步设计，减少临界区争用
- 数据库中主要的热点临界区：日志、锁、事务、SGA、字典缓存、…

Top Wait: BMSQL tpcc 30分钟统计采样

等待事件	等待时间（秒）	事件描述
log sync wait	13410.27	事务提交时等待日志写回日志文件中
log write buffer wait	2467.47	日志写入公共日志缓存时发生的等待
SGA page wait	2178.23	SGA中页面访问发生的等待
txn itl wait	1465.93	事务在行级锁上发生的等待
lock buffer wait	1092.09	访问公共锁表时发生的等待
SGA freelist wait	624.62	访问SGA页面空闲队列时发生的等待
useg global wait	210.20	绑定/解除绑定回滚段时发生的等待
txn slot wait	167.17	访问事务表时发生的等待
SGA hash wait	96.10	访问SGA Hash表时发生的等待
seg search cache wait	67.07	访问段分配缓存时发生的等待
…	…	…

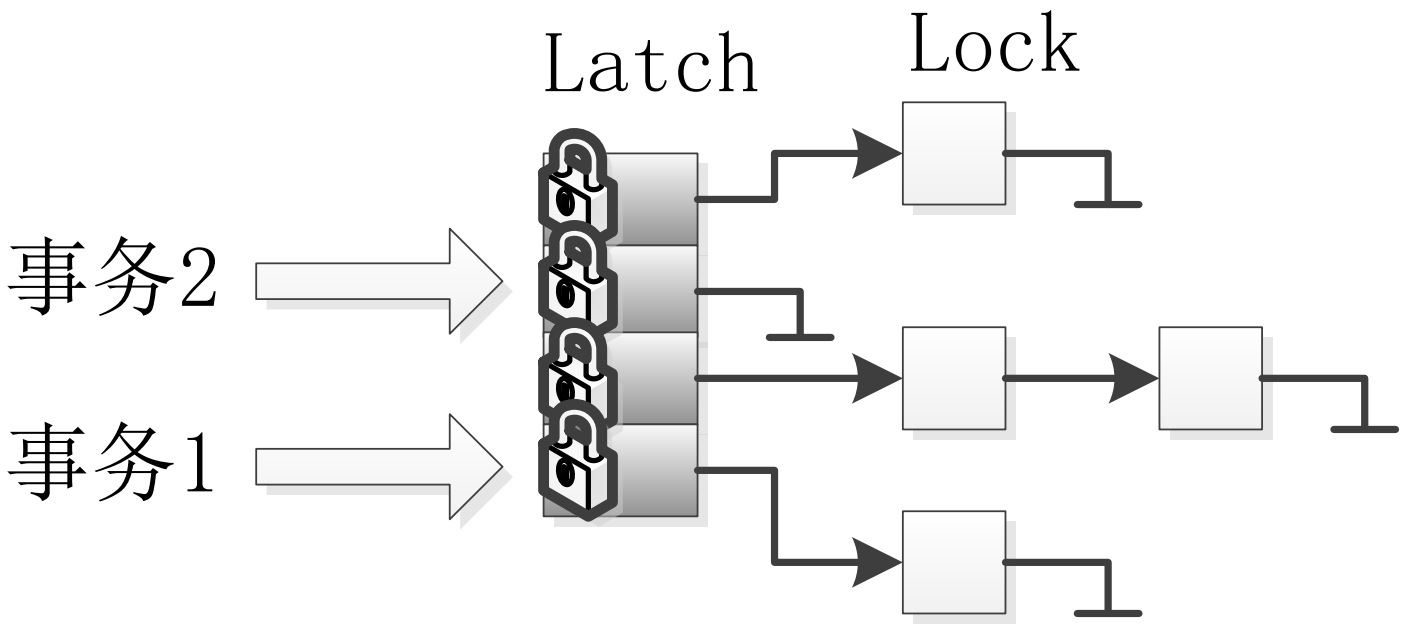
同步设计成为我们的主攻方向！

实战1：锁（Lock）

- 数据库中重要的结构，并发的热点
- 仅用于表级/字典锁（感谢MVCC！）
- Lock和Latch
  - Lock：逻辑概念，用于实现对逻辑对象（如表）的并发访问
  - Latch：物理概念，用于实现对临界区的同步访问

自然的优化

- 让访问不同锁资源的事务可以同时执行
- 基于HASH表的锁表设计



抽象：分区（Partition）

- **实现更细的访问粒度，并行的基础技术**
- **分区的种类**
  - 客体（静态）分区
  - 主体（静态）分区
  - 动态分区

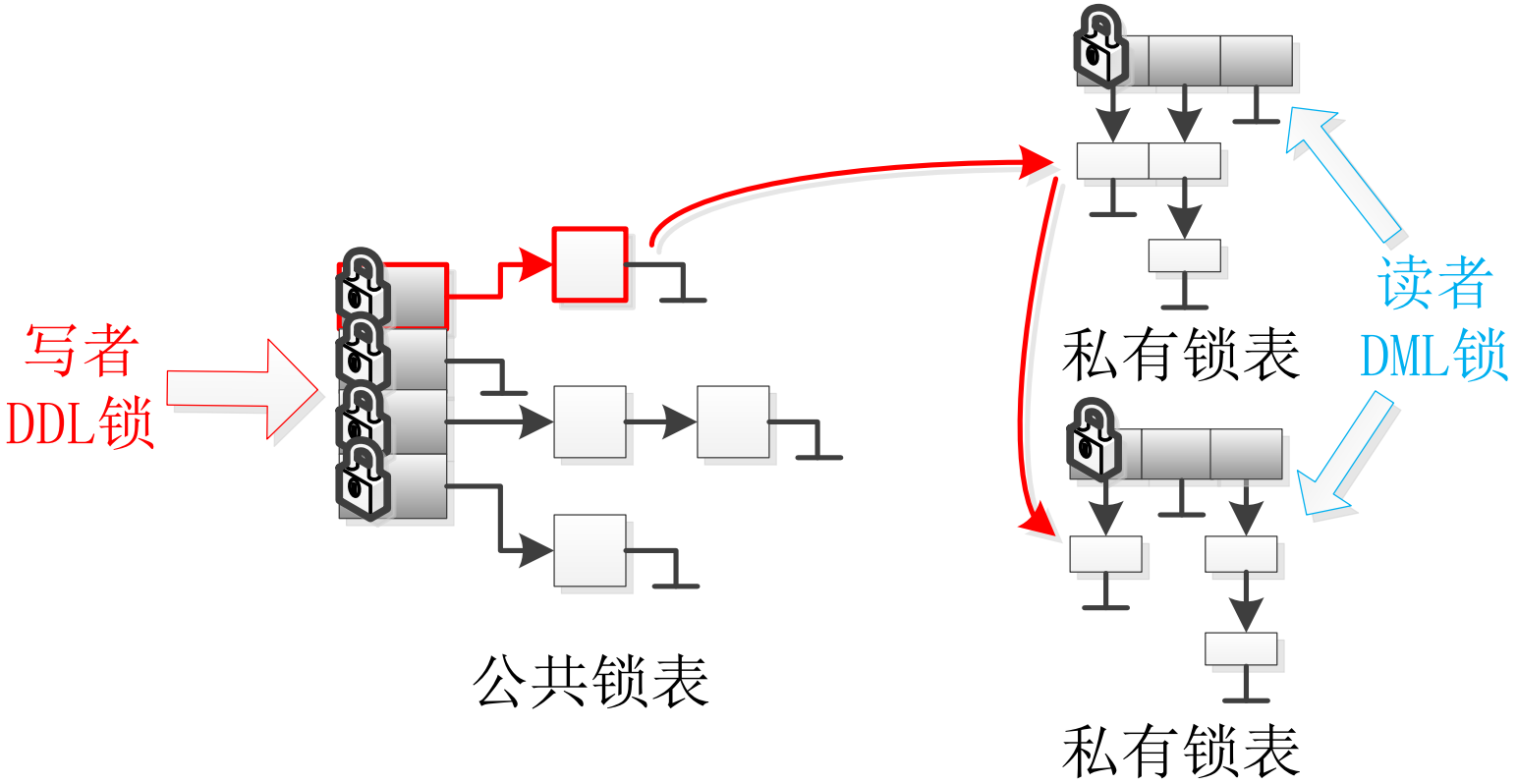
接下来...

- **如何应对访问热点？**
- **尝试对临界区（操作）进行分析：**
  - 大多数时候事务的加锁只是为了确保表不被删除或因为模式发生变化（小概率事件），可以据此分成两类：
  - 读者：DML（SELECT/INSERT/UPDATE/DELETE）操作
  - 写者：DDL（CREATE/ALTER/DROP TABLE/INDEX...）操作
- **优化思想：**
  - 将锁拆分成多个版本
  - 让读者和写者分别访问不同的版本（读写分离），以此来提高性能

抽象：复制（Replication）

- **基于版本，支持读写分离（MVCC or RCU?）**
- **牺牲写者，实现读者最短执行路径，适用于读多写少场景**
- **严格定义：**
  - 所有资源（如锁）被复制成一个全局版本和多个（每事务）私有版本
  - 根据访问特性将临界区（操作）拆分为读者和写者，满足：
  - 读者只存取（注意：不是读取！）私有版本，各读者并发执行，互不干扰
  - 写者严格串行，存取全局版本，并负责将其更新复制（同步）到受影响的所有私有版本中
  - 对于私有版本的访问需要同步（读者：单点争用 -> 多点并行）

最后：私有锁表



一个复杂些的例子

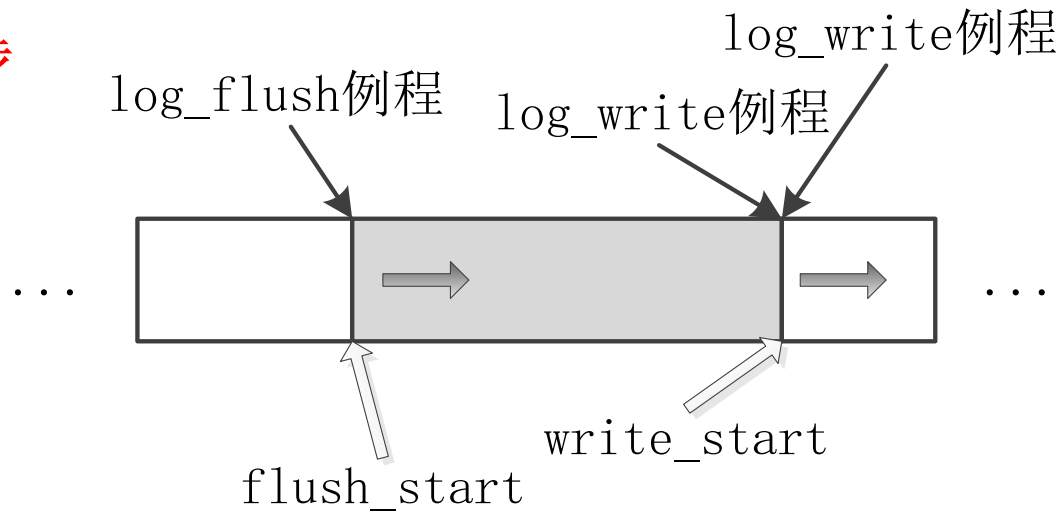
实战2：日志（Log）

- 缓存结构， OLTP数据库中最重要的同步热点
- 经典的非定长多生产者单消费者队列模型

初始实现：单一全局Latch同步

生产者（log_write）例程:
- 获得全局Latch
- 分配地址空间
- 拷贝日志
- 释放全局Latch

消费者（log_flush）例程:
- 获得全局Latch
- 计算要回写的地址空间
- 释放全局Latch
- 执行实际回写



公共日志缓存结构

如何分区？

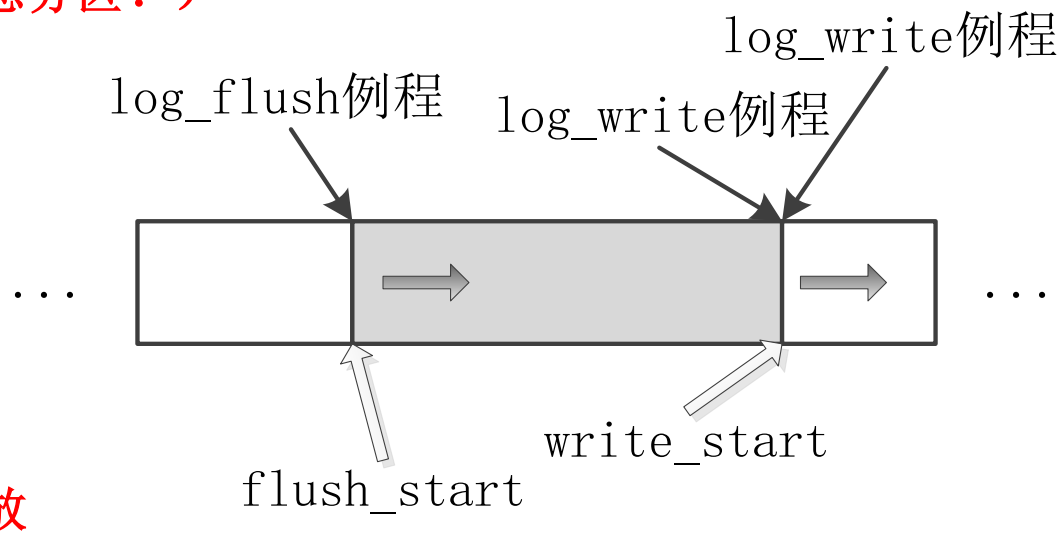
- **约束条件：**
  - 恢复协议约束：恢复动作是基于页面（Page-Oriented）的，同一页面上的恢复动作， 其执行顺序须与他们正常执行时的顺序相同（Repeating-History）（Page Partition?）
  - 持久性约束：事务在提交以前必须将其所有日志写回日志文件（Transaction Partition?）
- **换一种思路：**
  - 将整个写日志动作拆分成：分配地址空间+拷贝日志
  - 利用分区技术，实现多个拷贝日志动作的并行

基于动态分区的优化

改进实现：单一全局Latch拆分为单个Alloc Latch和多个Copy Latch

log_write例程:
- 获得Alloc Latch
- 分配地址空间
- 随机获得一Copy Latch（动态分区！）
- 释放Alloc Latch
- 拷贝日志
- 释放Copy Latch

log_flush例程:
- 获得Alloc Latch
- 计算要回写的地址空间
- 释放Alloc Latch
- 等待当前所有Copy Latch释放
- 执行实际回写

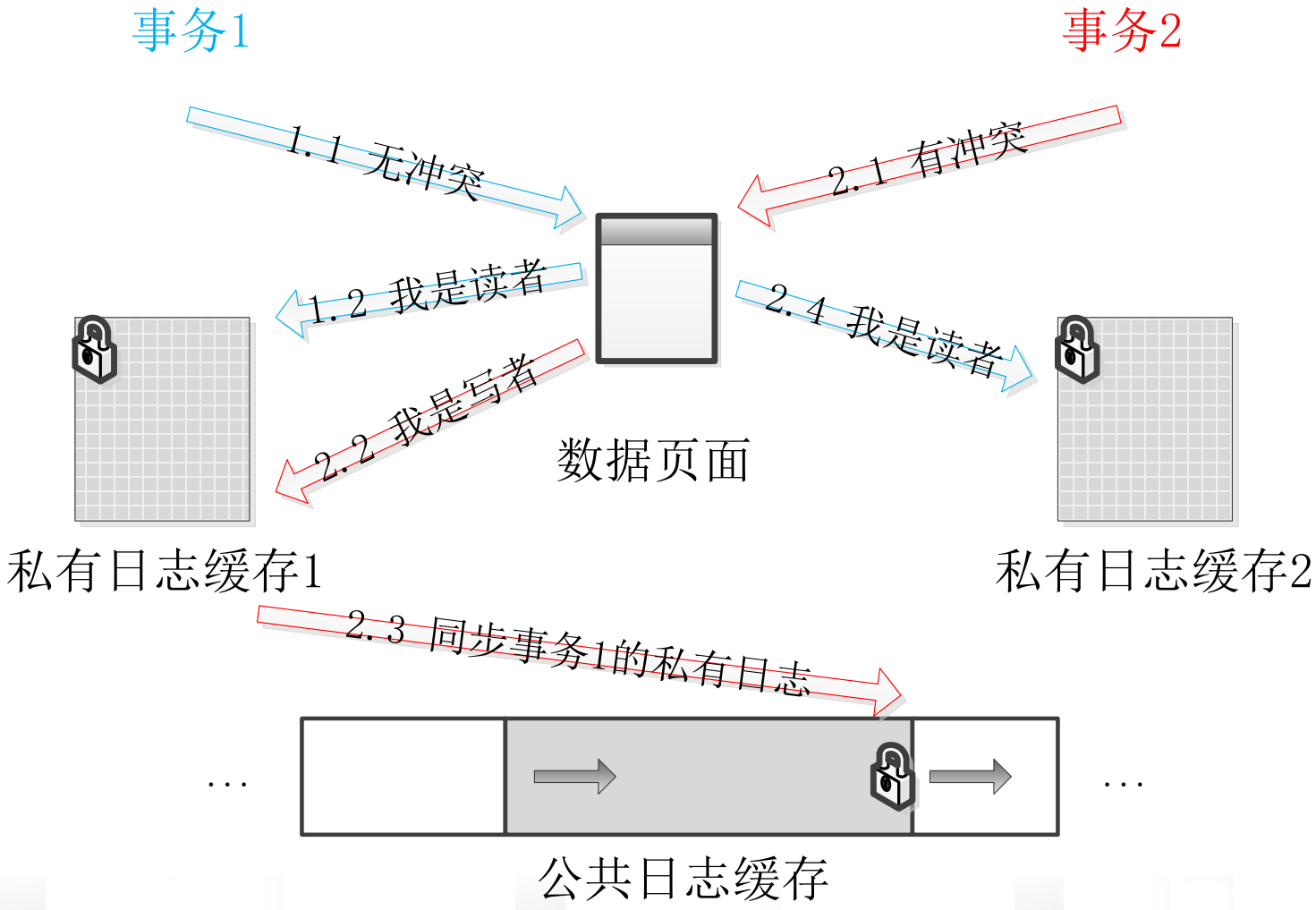


公共日志缓存结构

如何复制？

- **根据恢复协议，不同事务访问不同页面的恢复动作，其顺序是不相关的**
- **以页面为纽带，定义读者和写者：**
  - 何谓冲突？当事务在数据页面上产生日志时，如果有其他事务在该页面上保留有未写入公共日志缓存的日志，则发生冲突
  - 根据冲突定义读者和写者，写日志时，不冲突即为读者，否则即为写者

最后：私有日志缓存

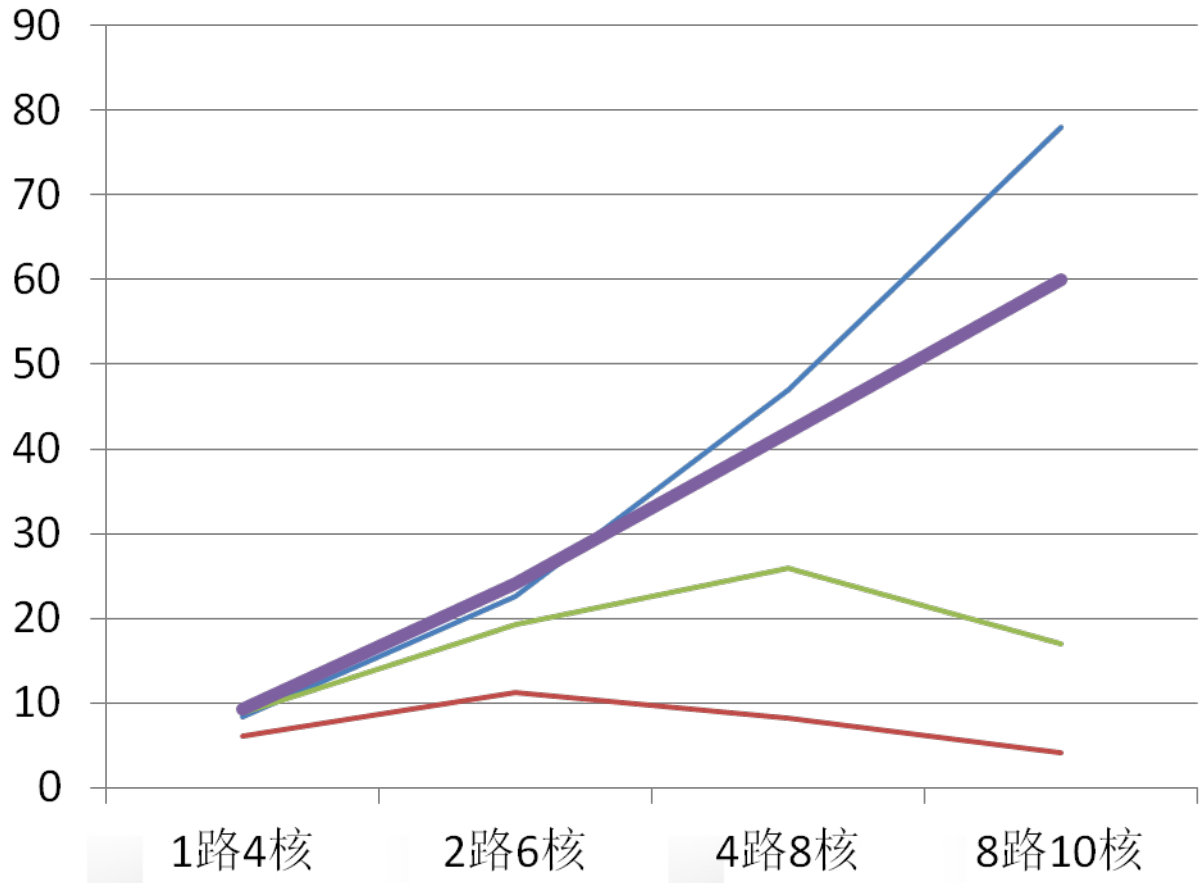

事务1
事务2
1.1 无冲突
2.1 有冲突
1.2 我是读者
2.4 我是读者
2.2 我是写者
数据页面
私有日志缓存1
私有日志缓存2
2.3 同步事务1的私有日志
...
...
公共日志缓存

部分工作一览

同步技术	典型临界区	统计
客体分区	锁/字典缓存/SGA…	16
主体分区	统计信息/堆内存…	12
动态分区	日志/事务表/List类…	24
复制	锁/日志…	6
其它临界区拆分技术	日志/段分配缓存/回滚段…	9
Latch扩展原语	字典缓存/SGA…	12
Atomic原语	List类/计数类…	5
…	…	…

我们的进展

单位：万笔交易/分钟



TPC-C
测试基准

BMSQL (SQL模式)
200w/200u

O*

MySQL InnoDB5.5

神通数据库
（上一版本）

神通数据库
（即将发布）

处理器

关于问题全景的思考

- **战略层面：让临界区变得更少**
  - 并发协议层（MVCC/…）
- **战术层面：让临界区变得更短**
  - 同步设计层（分区/复制/…）
- **战斗层面：让出入临界区变得更高效**
  - 原语实现层（Latch/Lock-Free/Memory Barriers/…）

神舟通用公司简介

公司概况

神舟通用数据技术有限公司致力于神通国产数据库产业化，隶属中国航天科技集团公司，是国内最具影响力的基础软件企业之一。公司提供神通数据库系列产品与服务，产品技术领先，先后获得30项数据库技术发明专利，在国产数据库行业处于领先位置。公司拥有北京研发中心、天津研发中心、杭州研发中心三家产品研发基地，与浙江大学、北航、北京大学、中科院软件所等高校和科研院所开展了深度合作，具有一大批五年以上的数据库核心研发人才。

主营业务

神舟通用公司主营业务主要包括神通关系型通用数据库、神通KStore海量数据管理系统、神通xCluster集群件、神通商业智能套件等系列产品研发和市场销售。基于产品组合，可形成支持交易处理、MPP数据库集群、数据分析与处理等解决方案。公司拥有40余名实战经验丰富的中高级数据库技术服务人员，可提供数据库系统调优和运维服务。公司客户主要覆盖政府、电信、能源、交通、网安、国防和军工等领域，率先实现国产数据库在电信行业的大规模商用。

公司发展历程

2012年

国家互联网应急中心批量应用

互联网几个大型工程相继上线，其中工程A管理数据量500TB以上，实现目前为止数据规模最大应用

2011年

电信行业国产数据库最大规模商用

中国联通全国集中综合结算系统全面采用神通数据库，实现国产数据库在电信行业的最大规模商用。

2008年

整合资源，成立专业公司

整合国产数据库技术和资源，成立神舟通用数据库专业公司，产品名称更改为“神通数据库”。

2004年

在航天和电子政务领域开展应用

发布OSCAR V5版本，在科技部组织的国产数据库测评中名列第一，在航天和电子政务领域开展应用。

2003年

获得863重点支持

以710所CAD/CAM中心为主体重组神舟软件公司，建立数据库事业部，获得科技部“十五”863计划重点支持。

1993年

联合浙大，开展技术攻关及产品研制

中国航天710所联合浙江大学，开展工程数据库管理系统（OSCAR）的技术攻关和产品研制工作。

主要产品

解决方案

神通BDC大数据解决方案

商业智能套件

神通K-Miner数据挖掘 | 神通K-Cuber OLAP系统

神通K-Front报表系统 | 神通K-Fusion ETL工具

集群套件

神通xCluster集群件

行业定制版本

神通KSTORE | 军用数据库

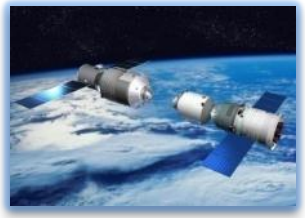
核心平台

神通数据库

市场应用









- 某运营商全国综合结算系统
- 无线网络优化系统
- 某省级运营商EDW系统
- 南方电网智能电站全网二次监控项目
- 公安部某局CA认证系统
- 公安部某局档案管理系统
- 国家碳排放交易系统
- 浙江某市质监局智慧数据分析系统
- 陕西省电子政务综合服务平台
- 山西省科技治超系统（省级、地市级）
- 广东省省级集中纳税服务平台
- 河南省环保厅上网电厂工况监控平台
- 山东省数字淄博项目
- 山东德州电子政务大集中项目
- 陕西省延安市“数字延安”项目
- 北京市农委农超一体化项目
- 天津档案馆档案管理系统
- 江西省监察厅交换平台系统
- 若干总装、总参、航天等国防军工行业信息化项目
- ……

讨论 & 致谢

联系我们: @神通数据库 @神通冯柯

www.shentongdata.com