

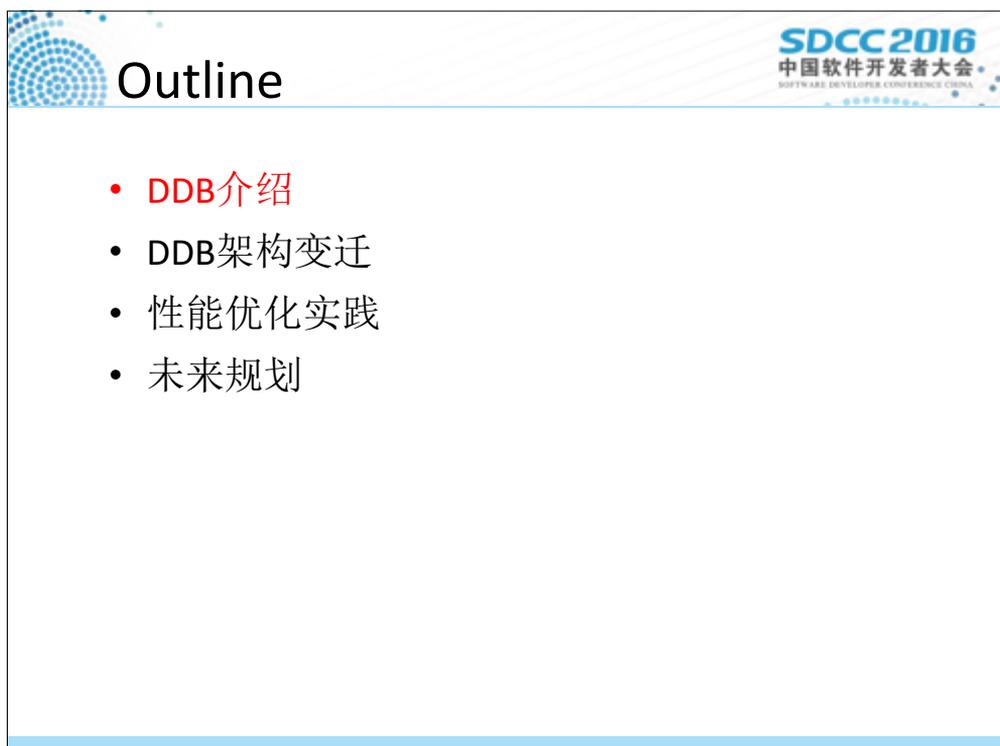


SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

网易DDB性能优化实践

嘉宾：马进（@马进app）
公众号：datamanagement

The slide features a light blue background with abstract patterns of dots and gears. The text is centered and uses a mix of blue and black colors.



Outline

- **DDB介绍**
- DDB架构变迁
- 性能优化实践
- 未来规划

The slide has a light blue header with the SDCC 2016 logo on the right and a blue footer bar at the bottom. The main content is a bulleted list.

DDB——十年一剑

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

• 分库分表数据库DDB

- 海量结构化数据存储，TB级别热点数据
- 高并发访问，应对OLTP在线事务型的应用
- 数据扩容，在线增删数据库节点，完善的DBA工具
- 透明分库分表，MySQL通信协议兼容

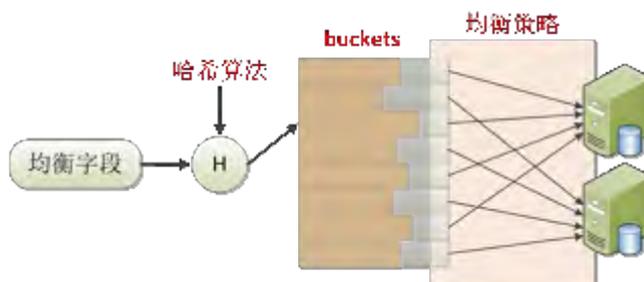


DDB功能价值

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

• 分区方案选型

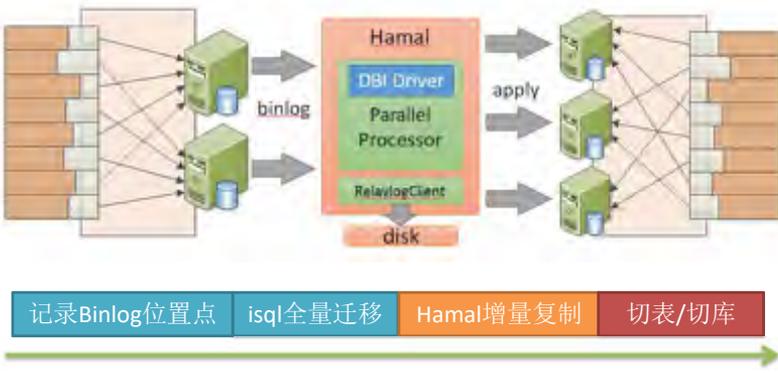
- 常规：取模哈希 + 桶（虚拟节点）
- 桶 + 取模哈希 + 均衡策略 = 单调性 + 均衡性 + 易用性
- 支持自定义哈希函数，可定制List和range分区
- 不能迁移的分区都是耍流氓



DDB功能价值



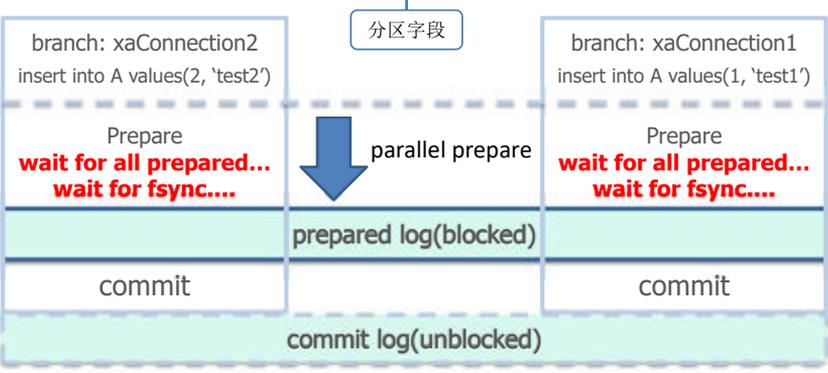
- 增量迁移工具Hamal
 - 库迁移场景：节点扩容扩容，集群机房迁移
 - 表迁移场景：表级扩容扩容，更改均衡策略/字段
 - 功能亮点：断点续传，并行复制



DDB功能价值



- 分布式事务
 - 透明的两阶段提交过程
 - 系统自动识别是否需要两阶段，优先一阶段
 - insert into A values(1, 'test1'),(2, 'test2');





DDB功能价值



- 分布式执行计划

```
mysql> explain select avg(age) from UserTest group by name limit 10,10;
+-----+
| PLAN
+-----+
| LIMIT/OFFSET
| This plan will be dynamically set disable/enable while running based on the underlying plan.
| ^
| /|\
| ||
| AGGREGATE
| Do:
| ^
| /|\
| ||
| PROJECT
| Project record to: SUM(age),COUNT(age),
| ^
| /|\
| ||
| GROUP
| Group By: name,
| ^
| /|\
| ||
| MERGE-SELECT
| SQL: SELECT SUM(age), COUNT(age), name FROM UserTest GROUP BY name ORDER BY name ASC
| Dist Node:
| dbn1[jdbc:mysql://10.120.146.129:3306/dbn1]
| dbn2[jdbc:mysql://10.120.146.129:3306/dbn2]
| dbn4[jdbc:mysql://10.120.146.130:3306/dbn4]
| dbn3[jdbc:mysql://10.120.146.130:3306/dbn3]
| Order by: name ASC, with merge sort.
+-----+
```



DDB功能价值



- 功能特性
 - 90%以上的SQL92兼容性，向MySQL语法靠拢
 - 命令行工具isql，管理工具DBAdmin
 - 支持数据节点手动或自动fail over
 - 基于hint的读写分离功能
 - 两种全局自增长ID实现
 - 更多丰富的hint功能
- 完善,无侵入的云端解决方案
 - 延展性: 查询服务器无状态，支持无缝水平扩展
 - 扩展性: 支持不同语言，不同实现的MySQL客户端访问DDB
 - 可用性: RDS数据节点基于IP漂移的高可用方案
 - 易用性: 云端一键部署，完善的WEB管理工具



Outline

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

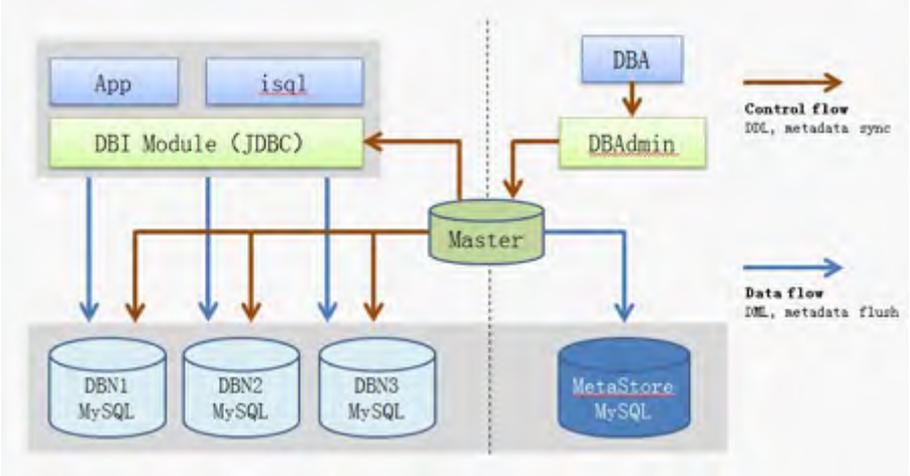
- DDB介绍
- **DDB架构变迁**
- 性能优化实践
- 未来规划



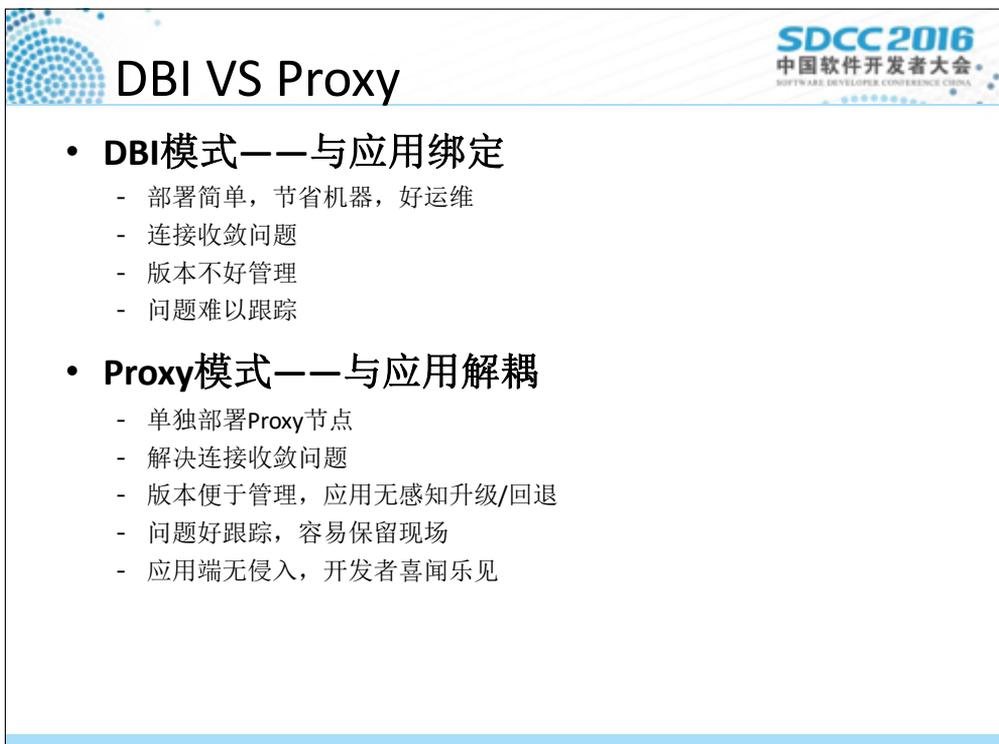
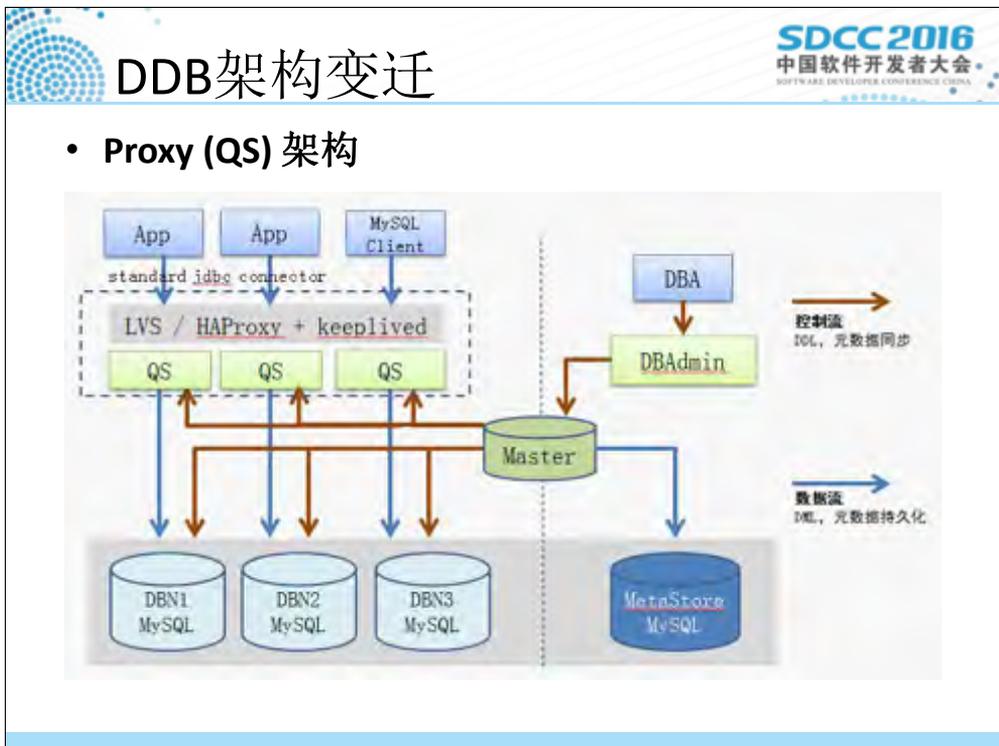
DDB架构变迁

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

- **DBI架构**



The diagram illustrates the DBI architecture. On the left, an application (App) and isql connect to a DBI Module (JDBC). The DBI Module connects to a Master MySQL instance, which in turn connects to three slave MySQL instances (DBN1, DBN2, DBN3). On the right, a DBA connects to a DBAdmin instance, which connects to the Master MySQL instance. The Master MySQL instance also connects to a MetaStore MySQL instance. Arrows indicate control flow (DBA to DBAdmin, DBAdmin to Master, Master to DBI Module) and data flow (App to DBI Module, DBI Module to Master, Master to DBN1-3, Master to MetaStore).



DDB架构变迁



SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

• DDB私有云服务

分布式数据库首页

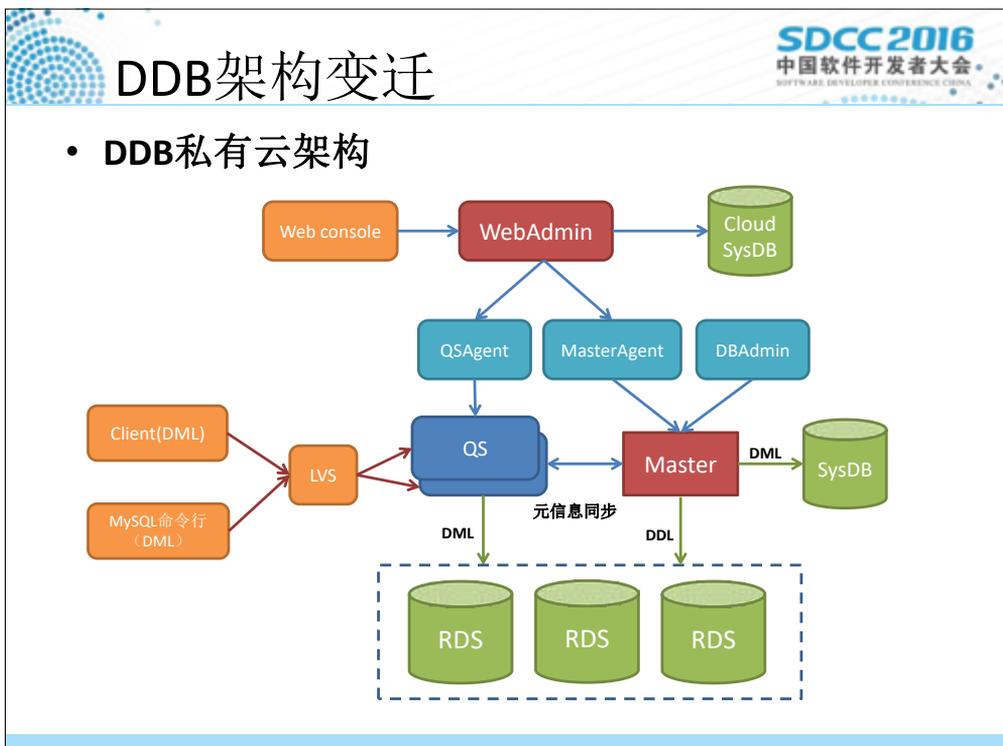
集群管理 > mjtest > SQL代理服务器管理

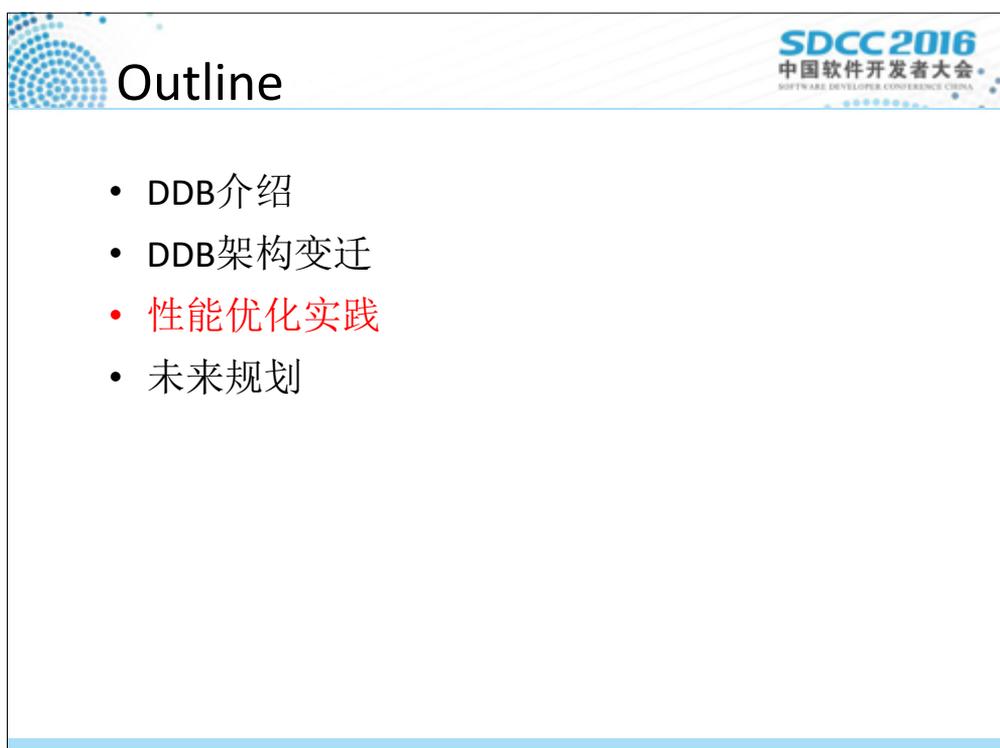
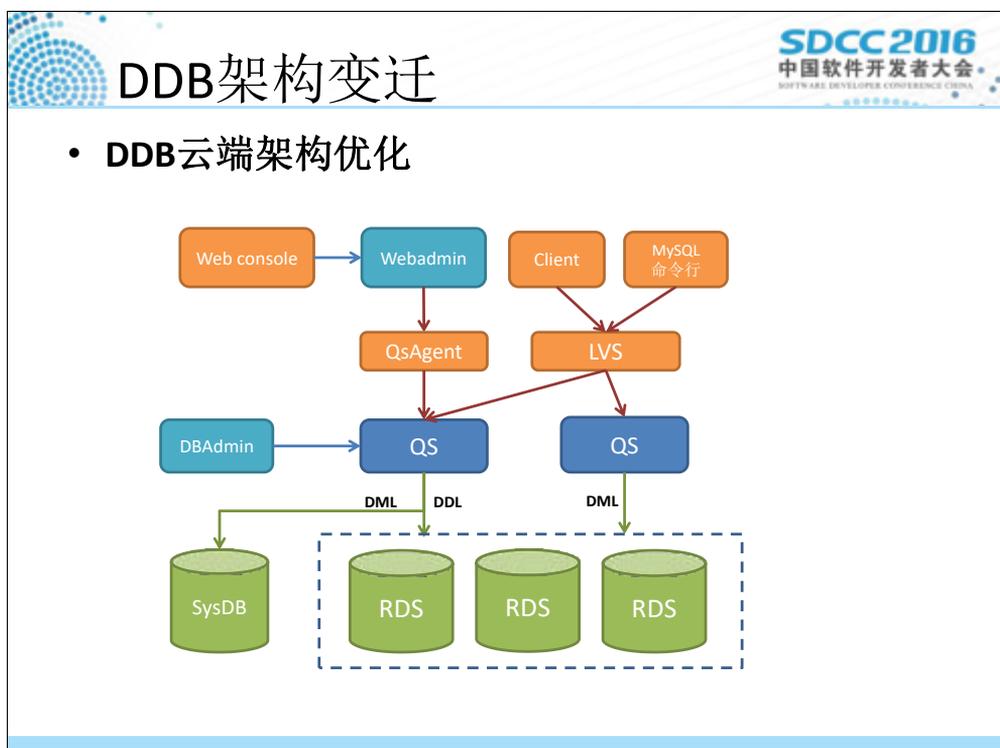
增加SQL代理服务器 删除SQL代理服务器 查看监控详情 查看连接状态 加锁表 还原备份 设置策略 下线 重连

ID	SQL代理服务器名称	规格	固定IP
01	mjtest-SQLPay-1	6GB内存, 4VCPU=4ECU	10.120.44.42:6001
02	mjtest-SQLPay-2	4GB内存, 4VCPU=4ECU	10.120.48.24:6001
03	mjtest-SQLPay-3	6GB内存, 4VCPU=4ECU	10.120.48.83:6001

集群管理

- 集群管理
 - hobtest
 - hctest
 - mjtest
- 用户管理
- 库管理
- 用户管理
- 数据库服务器管理
 - SQL代理服务器管理
 - 管理服务器管理
 - 备份管理
 - 操作日志
 - 维护任务管理
- 备份管理
- 操作日志





性能误区



- 分库分表比单机性能提升多少？
 - 分库分表benchmark一般不如单机，哪个好取决于MySQL是否到瓶颈
 - 分库分表性能指标：单节点策略下比吞吐率比单机折损多少（DDB在18%–35%）

DDB Proxy

解码
编码

语法解析

执行计划

多机数据

DBN1

DBN2

MySQL Server

解码
编码

语法解析

执行计划

单机数据

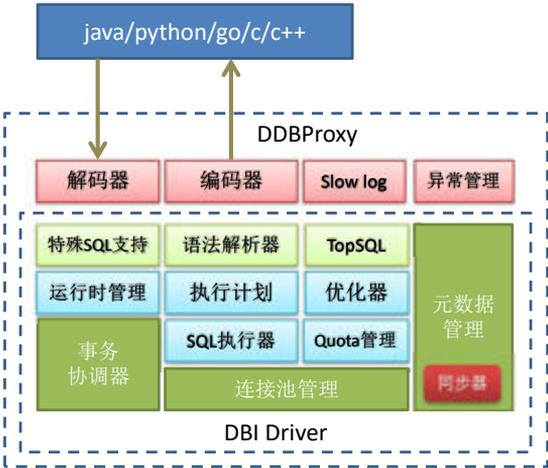
TB1

TB2

Proxy Buffer优化



- DDBProxy模块实现**
 - 基于DBI做分库分表
 - Server层做编码解码
 - 特殊SQL支持 (show命令)
 - 基于netty4网络框架构建
- 结果集编码**
 - 结果集大小不可知
 - 编码以列值为单元
- 存在问题**
 - Buffer碎片化
 - NIO线程切换过多

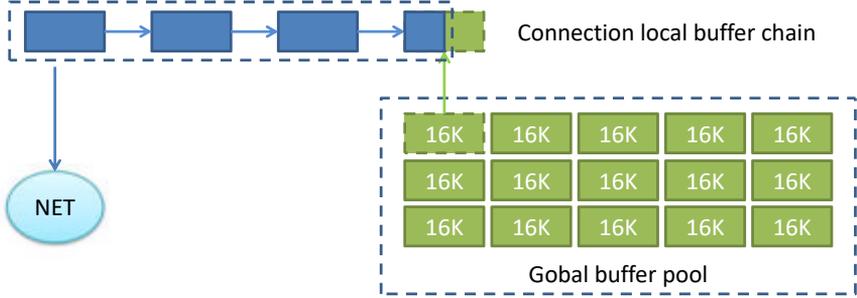


The diagram illustrates the architecture of DDBProxy and its interaction with the DBI Driver. At the top, a blue box labeled 'java/python/go/c/c++' represents the application layer. Below it, a dashed box labeled 'DDBProxy' contains several components: a '解码器' (Decoder) and '编码器' (Encoder) for data exchange; 'Slow log' and '异常管理' (Exception Management) for monitoring; and a 'DBI Driver' section which includes '特殊SQL支持' (Special SQL Support), '语法解析器' (Parser), 'TopSQL', '运行时管理' (Runtime Management), '执行计划' (Execution Plan), '优化器' (Optimizer), '事务协调器' (Transaction Coordinator), 'SQL执行器' (SQL Executor), 'Quota管理' (Quota Management), '连接池管理' (Connection Pool Management), and '元数据管理' (Metadata Management) with a '同步器' (Synchronizer) at the bottom right.

Proxy Buffer优化



- 优化思路
 - 所有Buffer以16K的大包为单位写入网络
 - 每100行 (why?) 列值的编码合并到16K的Buffer中
 - Buffer编码溢出后申请另一个16K的Buffer串联起来
 - Global buffer pool + Connection local buffer chain

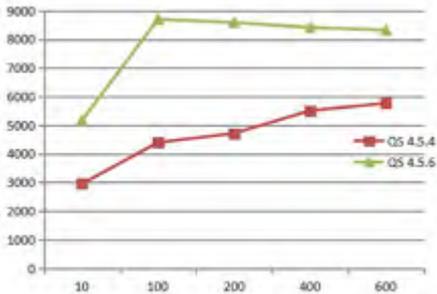


Proxy Buffer优化



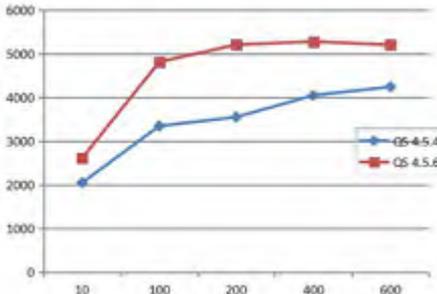
- 优化效果

SQL: select * from Blog where ID = ?
数据大小: 20字符 (小数据结果集)



Iteration	QS 4.5.4 (Time)	QS 4.5.6 (Time)
10	~3000	~5000
100	~4500	~8800
200	~4800	~8500
400	~5500	~8300
600	~5800	~8200

SQL: select * from Blog where ID = ?
数据大小: 4K字符 (大数据结果集)

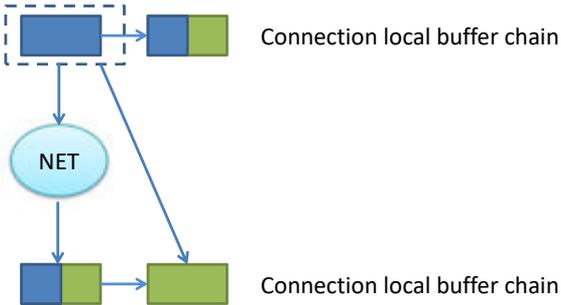


Iteration	QS 4.5.4 (Time)	QS 4.5.6 (Time)
10	~2000	~2500
100	~3500	~4800
200	~3800	~5200
400	~4200	~5300
600	~4300	~5200

Proxy Buffer优化



- 进一步优化
 - Buffer溢出后中断编码，当前Buffer写入网络后继续
 - 判断溢出在**一行**数据编码后，写完的Buffer挂到最后面
 - 一般BufferList不会超过2个（除非一个数据行超过16K）



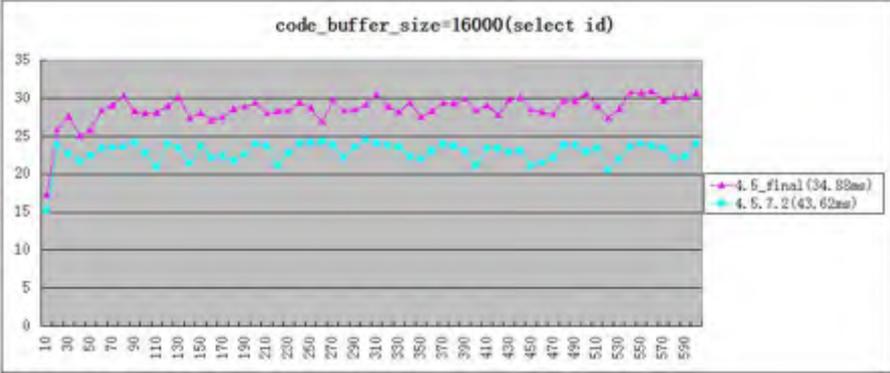
The diagram illustrates the data flow in a proxy buffer optimization scenario. It shows two 'Connection local buffer chain' components, each represented by a blue and green bar. A dashed box encloses the first buffer chain. An arrow points from this box to a central circle labeled 'NET'. From the 'NET' circle, an arrow points to the second buffer chain. Additionally, a direct arrow points from the first buffer chain to the second buffer chain, bypassing the network. This represents the state where a buffer is written to the network and then the next buffer is immediately available for writing.

Proxy Buffer优化



- 进一步优化效果

select id from sbtest1 limit 1000



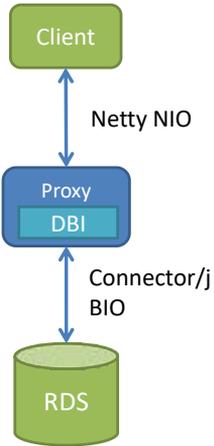
The graph shows the execution time in milliseconds for a query with a code buffer size of 16000. The x-axis represents the number of rows (from 10 to 590), and the y-axis represents time in milliseconds (from 0 to 35). Two data series are plotted: a purple line for '4.5_final (34.88ms)' and a cyan line for '4.5.7.2 (43.62ms)'. The purple line consistently stays below the cyan line, indicating better performance. The purple line fluctuates between approximately 25ms and 30ms, while the cyan line fluctuates between approximately 20ms and 25ms.

DBI NIO优化



- 通信模式
 - Proxy <—> Client: Netty NIO
 - Proxy <—> DBN: Connector/j BIO
 - BIO阻塞式通信，占用大量CPU

- 优化方法
 - 思路1：自研到MySQL的NIO驱动
 - 思路2：将Connector/j的底层通信依赖Netty
 - 优化效果：CPU使用率1.2 – 4倍提升



```

graph TD
    Client[Client] <-->|Netty NIO| Proxy[Proxy  
DBI]
    Proxy <-->|Connector/j  
BIO| RDS[(RDS)]
      
```

OSC



- OSC解决方案
 - DDB & pt-osc：基于锁的全量和触发器的增量
 - 触发器导致线上事务变慢，锁释放变慢
 - 全量扫描加锁导致锁冲突加剧
 - 线上实施容易大量**锁超时**

- 优化方法
 - 基于binlog实现增量更改
 - 全量不加锁，增量用replace幂等语义
 - 实现：DDB HamalSet & gh-ost
 - 优势：对线上几乎**零影响**

Outline

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

- DDB介绍
- DDB架构变迁
- 性能优化实践
- 未来规划

未来规划

SDCC 2016
中国软件开发者大会
SOFTWARE DEVELOPER CONFERENCE CHINA

- 架构优化
 - 优化管理架构，插拔所有平台
 - 通用平台运维工具DDBAdmin
 - 数据迁移服务抽离和分治
- 蜂巢DDB

镜像仓库



负载均衡



关系型数据库



对象存储



缓存



 - 云端一键部署，全方位管家服务
 - 高可用，高可靠，低成本
 - 2017年上半年，敬请期待

