

SACC

# 2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

架构设计 · 自动化运维 · 云计算

## 低成本和高性能 MySQL云架构探索

## 关于我

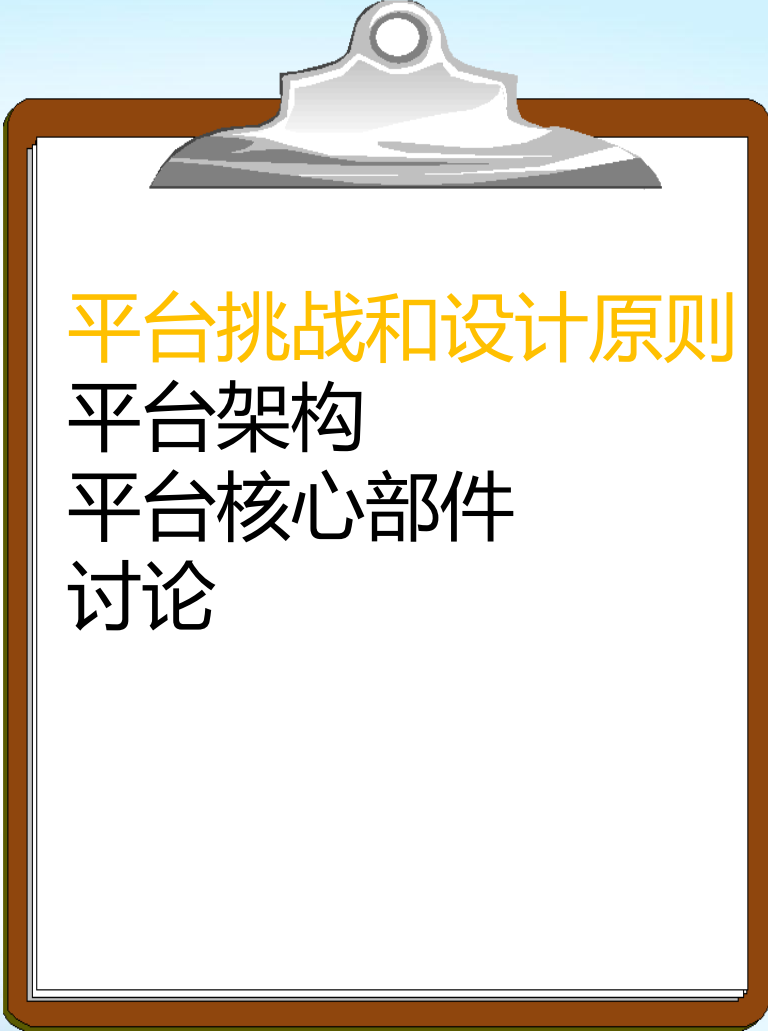
- 淘宝核心系统资深技术专家 余锋
- 超过15年互联网行业的网络、内核以及底层软件开发经验
- 专注于高性能分布式服务器的研究和实现
- 擅长构建大规模集群服务器
- 对数据库系统和分布式文件存储有深入的研究

SACC

2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

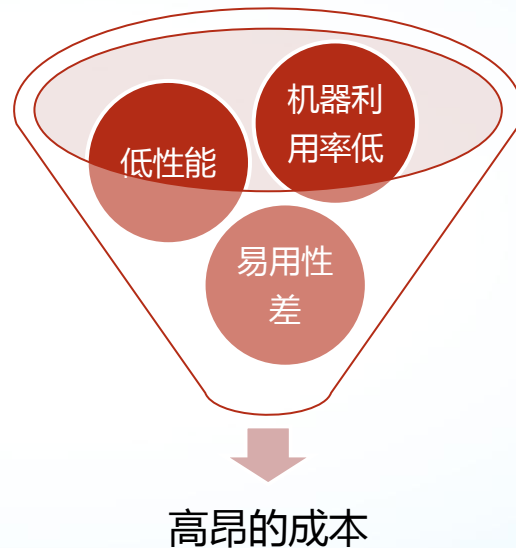
架构设计 · 自动化运维 · 云计算



平台挑战和设计原则  
平台架构  
平台核心部件  
讨论

# MySQL运维面临的问题

- 易用性偏差
- 性能
  - 软硬件未经优化，处于原始阶段
  - 不同阶段的软硬件性能相差巨大
- 集群
  - 主备不完全同步和备机利用率低
  - 以业务划分，集群分散管理，运维成本高



## MySQL平台化要解决什么？

- 高效MySQL数据库服务支持，自动化运维
- 提高资源利用率，降低平台成本
- 长期运行，对外屏蔽软硬件的变化



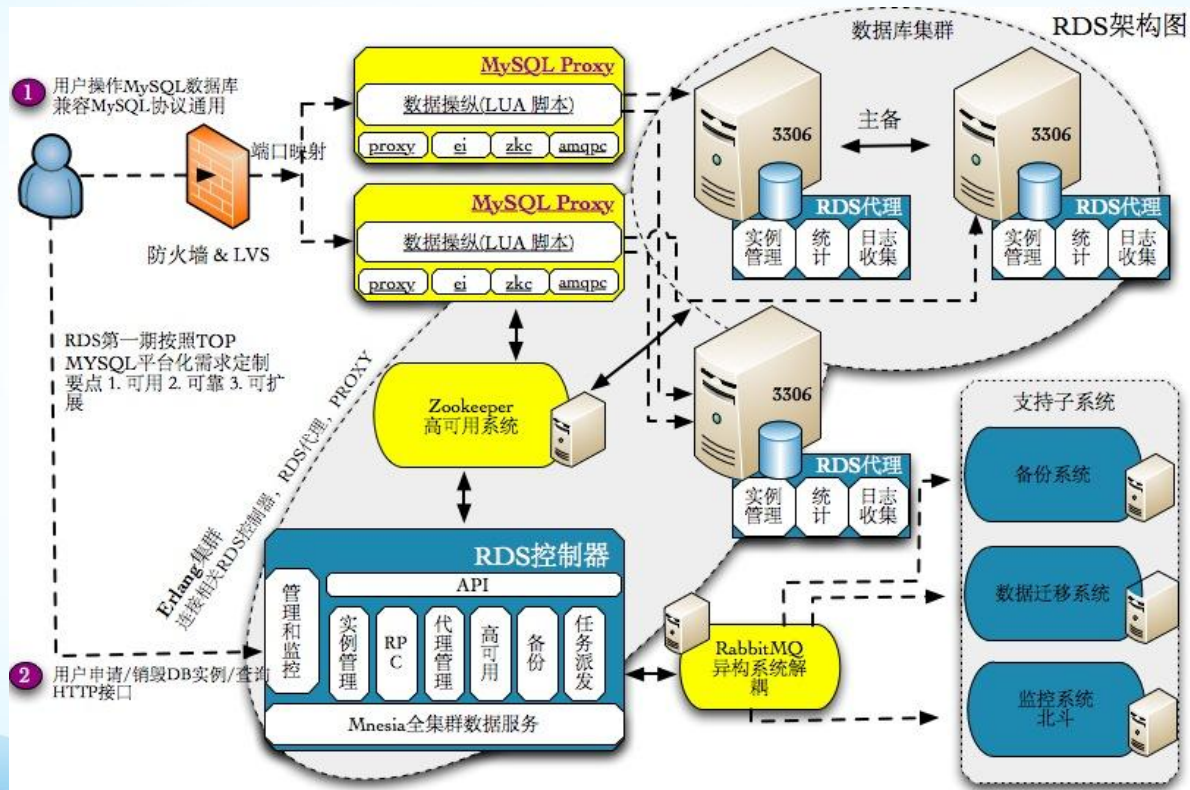
## MySQL平台设计原则

- 平台对外保持单一入口，对内维护单一的资源池。
- 保证服务的高可用性，消除单点故障
- 保证系统是弹性可伸缩的，可以动态的增加、删减计算与存储节点。
- 保证分配给用户的资源也是弹性可伸缩的，资源之间相互隔离。



平台挑战和设计原则  
平台架构  
平台核心部件  
讨论

# 平台架构第一版



SACC

2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

架构设计 · 自动化运维 · 云计算



# 第一版的经验和教训

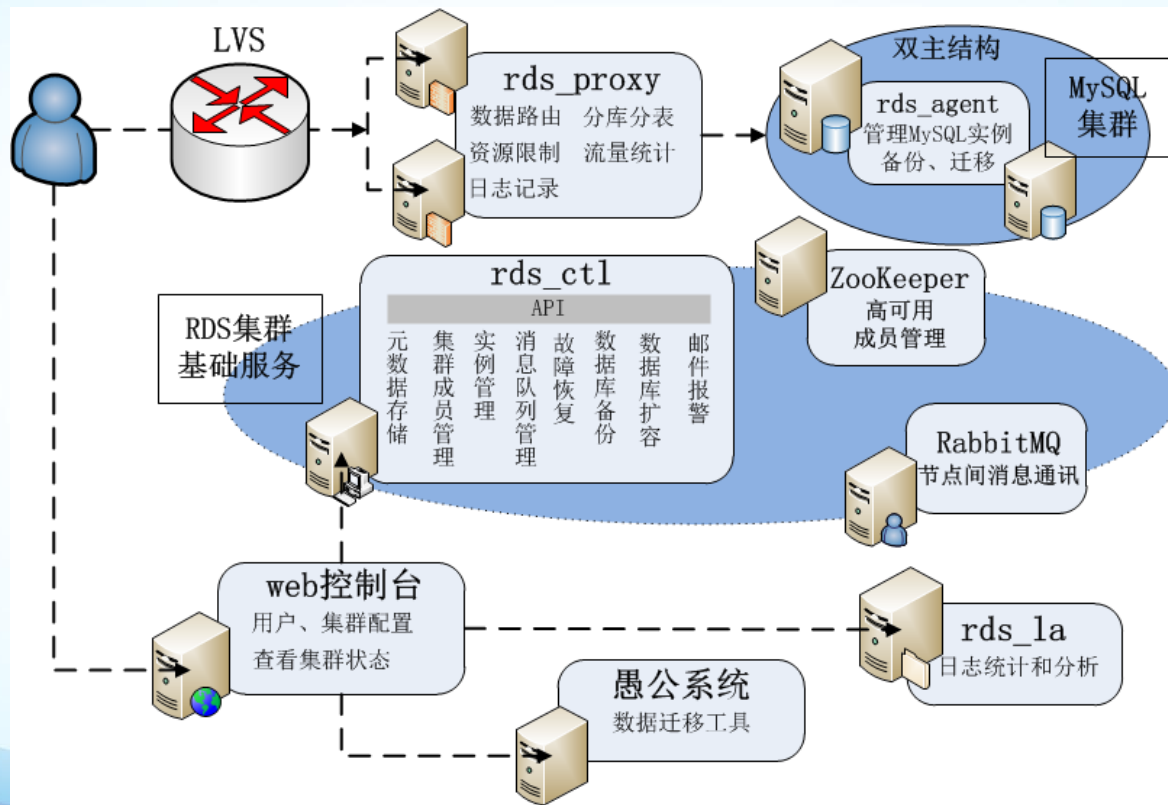
## ■ 经验

- 采用开放成熟的第三方部件的好处
- 开放的平台，方便用户扩展
- 热部署和升级对不停机维护的意义
- 容错系统设计的重要性

## ■ 教训

- 保持和MySQL的绝对兼容的重要性
- 数据访问主路径必须短且稳定
- proxy 性能、稳定性和成本的关系
- 日志实时收集和处理的难度
- 系统各部件部署的粒度，减少系统升级带来不良影响

# 平台架构第二版

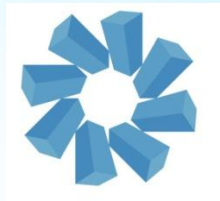


## 平台特性

- 平台足够稳定，支持热升级
- 支持几千台物理机规模，
- 以对用户透明的形式提供主从热备、数据备份、迁移、容灾、读写分离、分库分表功能
- 资源隔离，按需分配和限制CPU、内存和IO资源
- 不影响提供数据服务的前提下根据用户业务的发展动态的扩容和缩容
- 屏蔽数据节点不同的软硬件差异

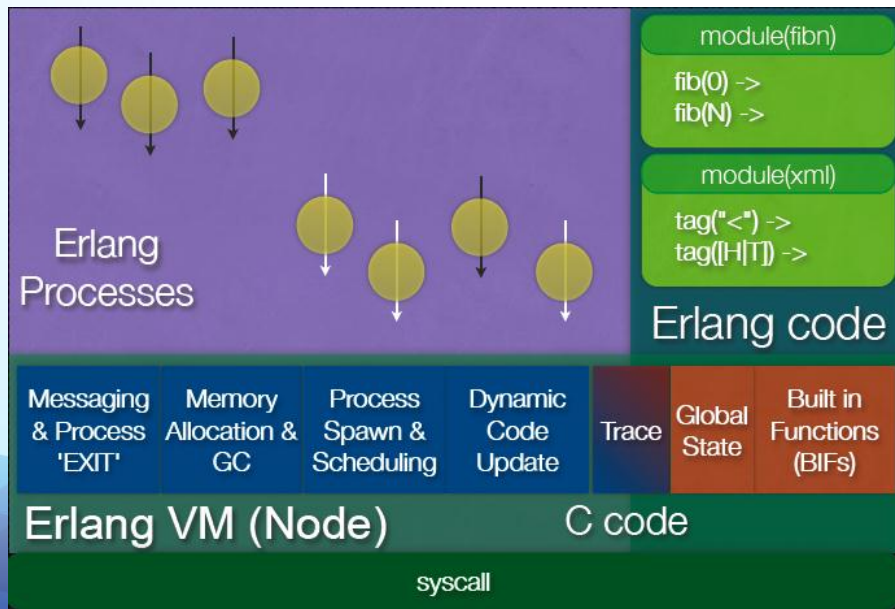
## 平台概况

- 稳定性生产系统验证过
- 依赖的开源组件：Mnesia、Lvs、RabbitMQ、ZooKeeper
- 代码规模
  - 核心使用以高性能、健壮以及可伸缩性出名的Erlang语言开发
  - 5万行Erlang代码，3万c代码，2万其他代码
  - 六人团队，历时1年



# 为什么要用Erlang实现

- 高并发,高性能,集群易扩展
- 时间检验的高可靠
- 强大的管理功能,方便的问题定位支持
- 强大的交互性, 与其他系统整合能力



SACC

2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

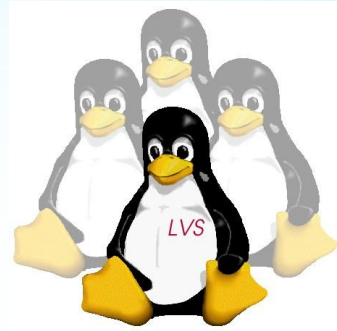
架构设计 · 自动化运维 · 云计算



平台挑战和设计原则  
平台架构  
平台核心部件  
讨论

# LVS

- 安全隔离
- L4流量切割，实现负载均衡
- Proxy故障自动转移



# ZooKeeper

- 作为配置服务器
- 提供分布式锁
- MySQL插件监控所有实例的可用性





# RabbitMQ

- Erlang消息机制与AMQP极度吻合
- 系统中各节点间的可靠通信（不包括SQL查询、日志等大数据流的传输）
- 标准的AMQP协议，方便第三方对接和扩展
- 工作流方式消息流动，方便监控



## 角色描述

- Controller服务器
- Proxy服务器
- Agent服务器
- API/Web服务器
- 日志分析服务器
- 信息统计服务器
- ...

## Mnesia分布式数据库

- Mnesia支持事务
- 无中心点
- 内存数据库，软实时

## 分库分表

- 对用户透明的分库分表 ( shard / horizontal partition ) 功能

# 容灾

- 对用户透明的故障恢复过程

SACC

2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

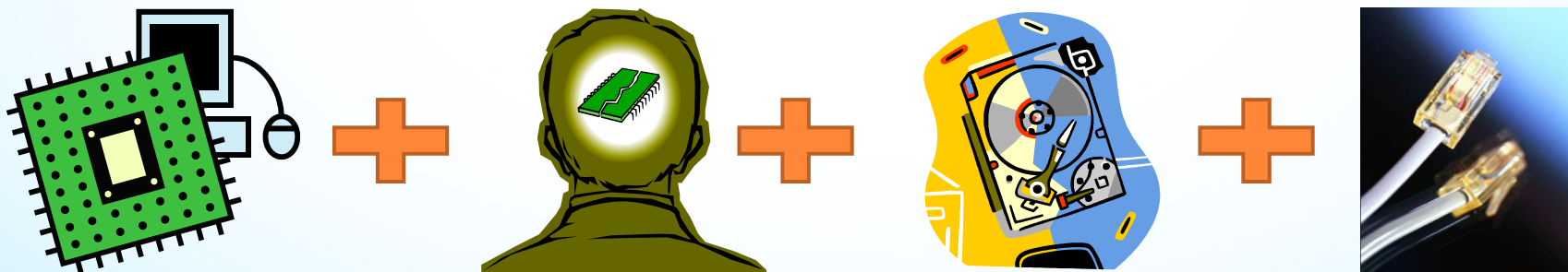
架构设计 · 自动化运维 · 云计算

## 读写分离

- 提高从库的利用率
- 时间维度？

## 资源隔离

- 通过CGroup的cpuset、memcg以及blkio子模块分别限制用户的MySQL进程最大可以使用的CPU使用率、内存和IOPS
- 对用户SQL执行过程中索引使用情况、IO操作数量等进行分析，指导Proxy增加延迟的方法去限制用户的QPS，达到了减小该用户消耗的系统资源的目的



## 资源调度

- 用户级别的资源预留??
- 系统自动迁移??

小实例

N : 1

中等规模

1:1

大规模

1:N

SACC

2012中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2012

架构设计 · 自动化运维 · 云计算



## 数据安全

- 支持SSL连接
- 通过白名单来设置允许访问数据库的IP地址列表
- Proxy会把用户所有的数据库操作记录到日志分析服务器，扫描检查安全漏洞
- Proxy可以根据安全部门的要求拦截各种类型的SQL语句
- 误操作删除数据又没有备份的情况可以通过Flashback工具恢复数据
- ...

## 下一阶段重点

- 提高各部件特别是proxy的效率，降低平台自身的成本消耗
- 资源隔离和调度完善, 进一步降低成本
- 和已有系统的融合，提高易用性

# Q&A

<http://yufeng.info>

新浪微博：@淘宝褚霸

**SACC**

**2012中国系统架构师大会**

SYSTEM ARCHITECT CONFERENCE CHINA 2012

架构设计 · 自动化运维 · 云计算