网易云 beta
cloud.netease.com

开源云端数据库架构

网易 杭州研究院
后台技术中心 郭忆
新浪微博：@郭忆_宝

目录

- 网易云数据库
- 系统架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

产品方使用MySQL遇到的问题？

- 硬件采购周期长，沟通协调成本高，数据库部署的需求难以得到快速响应。
- 硬件资源利用率低，难以做到按需使用，弹性扩容。
- 服务可用性差，数据可靠性难以保证。
- 运维自动化程度低，人力成本高。
- 监控报警不够完善，出现问题缺少系统诊断方法。

高额的成本+差的用户体验

云计算

- 云计算：

- 云平台的三个层次：

IaaS	• 计算、网络、存储 • 系统管理员
PaaS	• 语言运行环境、数据库、对象存储 • 程序开发人员
SaaS	• 邮箱、搜索、相册 • 最终用户

DBaaS

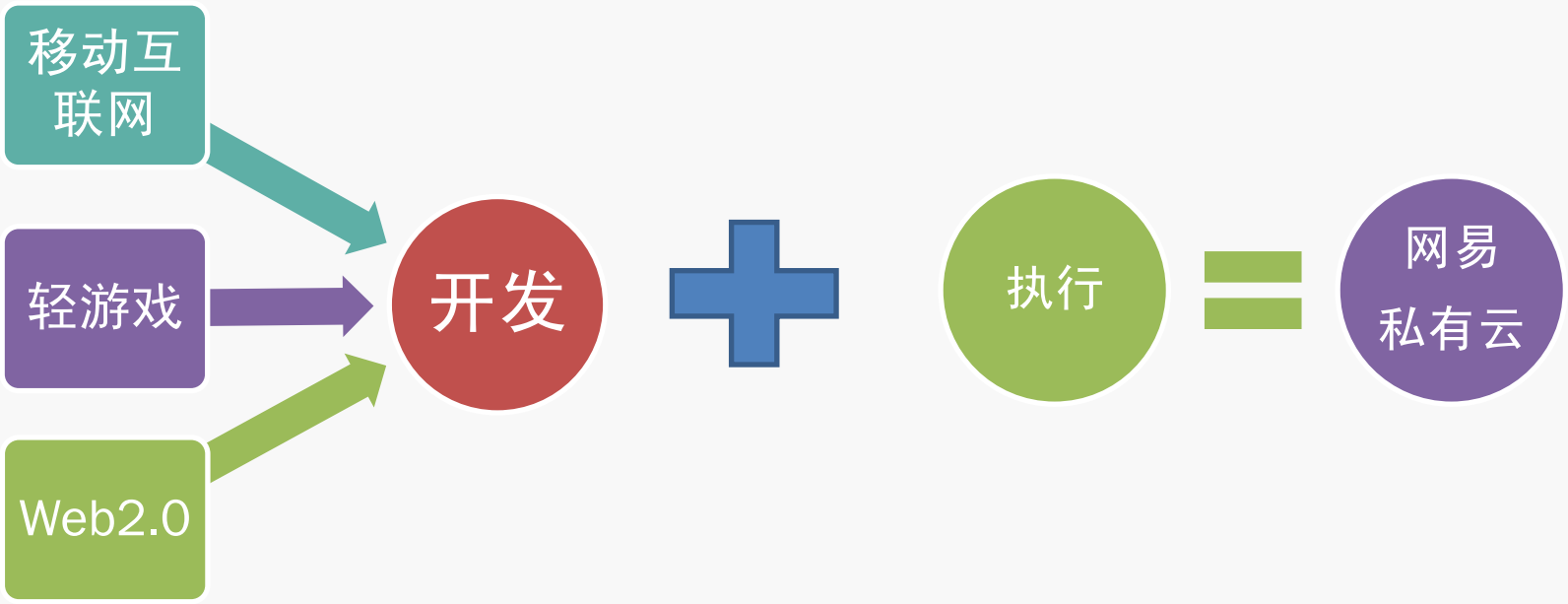
DBaaS(DataBase as a Service)，数据库即服务，云环境下的数据库托管平台，将数据库本身作为一种云端资源，以服务的方式提供给应用开发人员。

主要优势：

云计算

- 云计算不是天生为数据库设计的，数据库“云”化过程中，存在以下挑战：

网易私有云环境



網易 网易云 beta cloud.netease.com

网易私有云环境

网易私有云环境

系统特色

重视服务质量

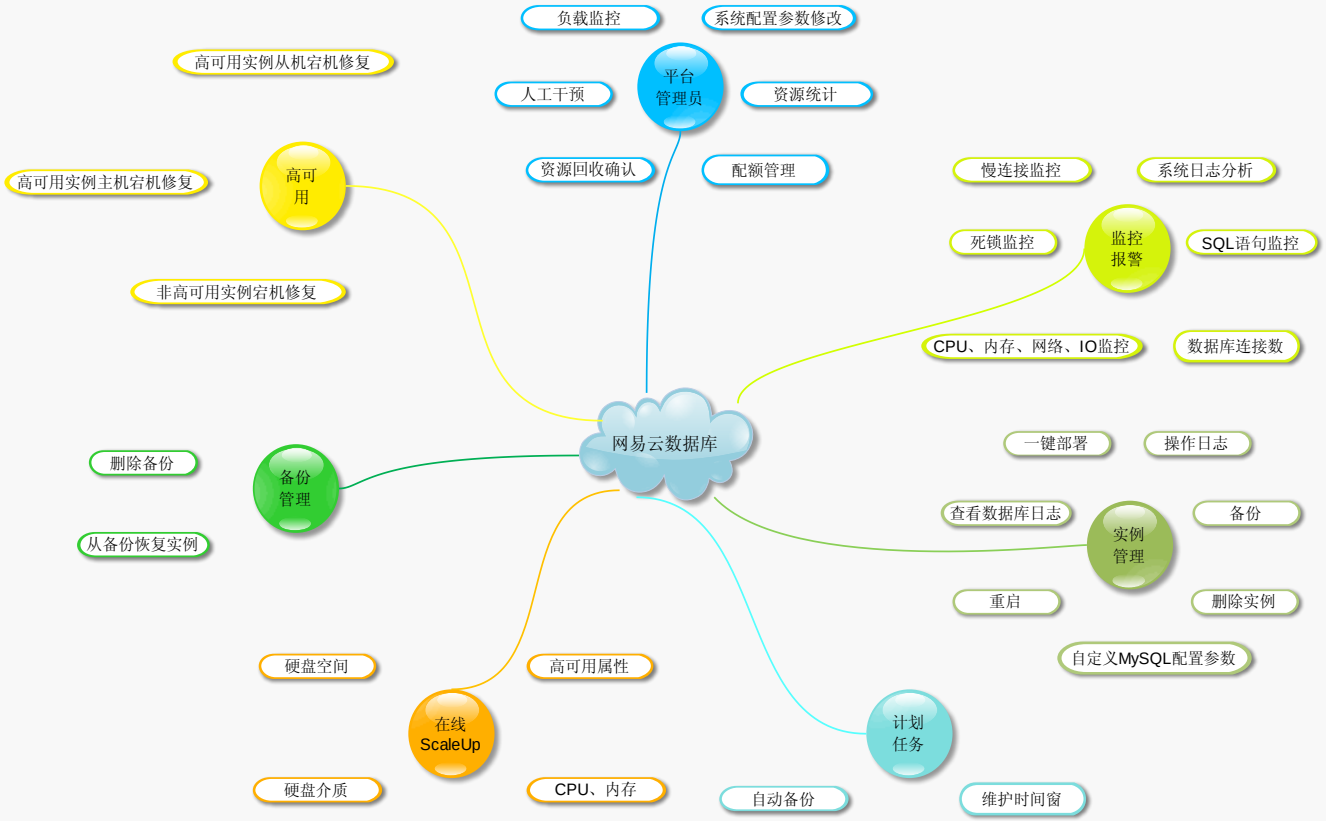
- 拒绝超售
- 高可用&高可靠

重视选择和控制能力

- 自定义CPU和内存组合
- 开放配置参数修改

网易私有云环境

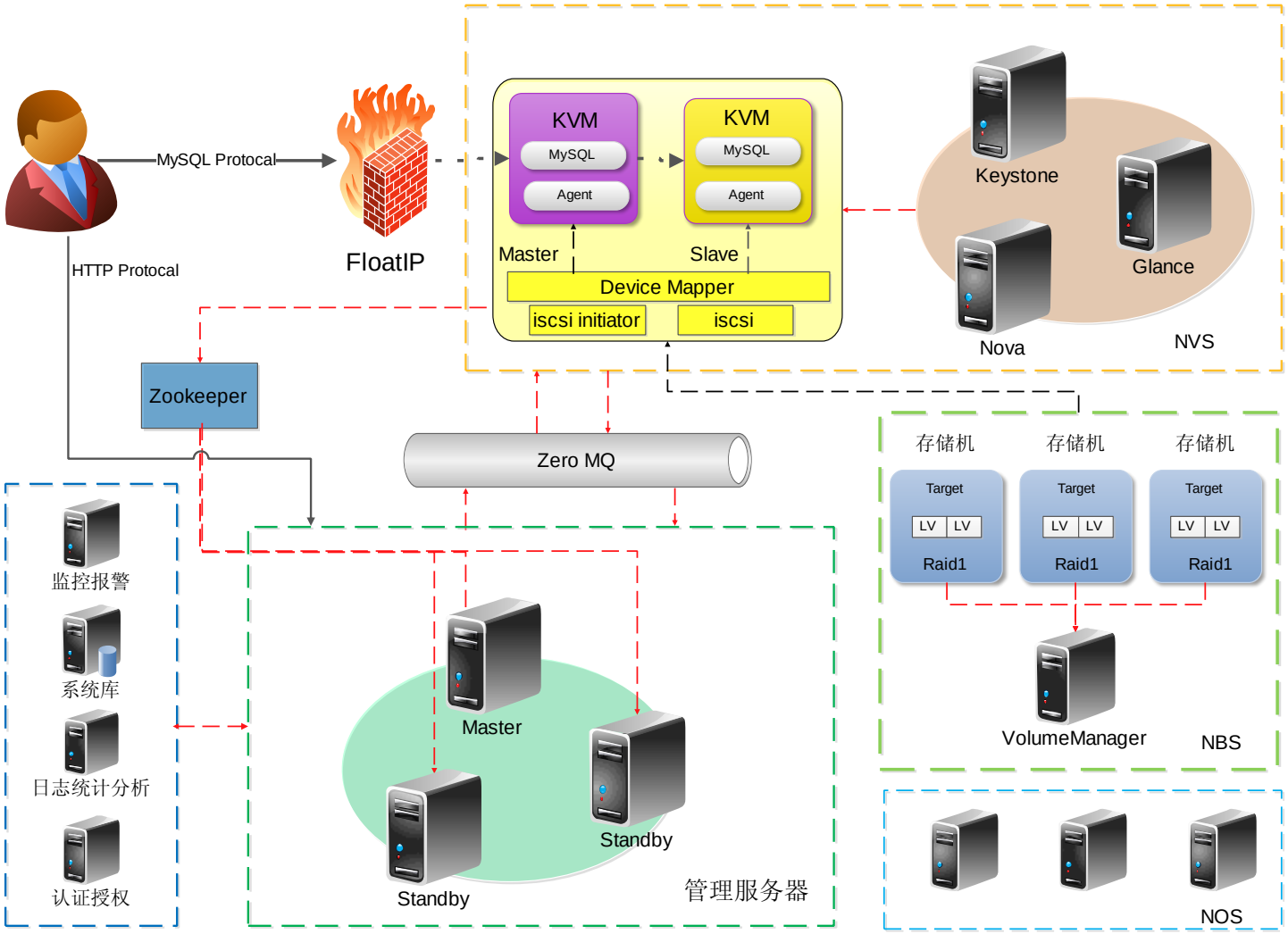
- Netease RDS是由网易数据库技术团队基于开源的MySQL打造的关系数据库云托管平台，构建于网易私有云IaaS服务之上，面向网易众多的互联网和移动终端产品，旨在解决当前数据库管理遇到的诸多问题。



目录

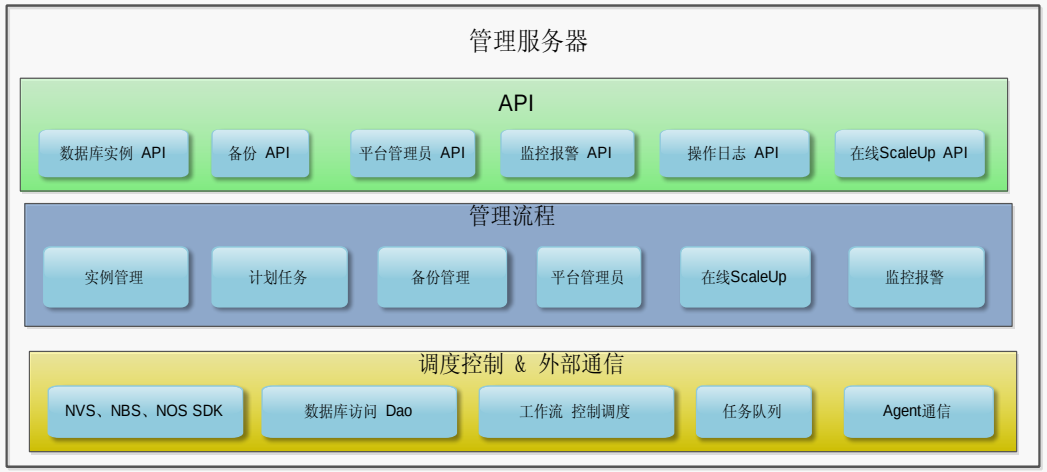
- 网易云数据库
- 系统架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

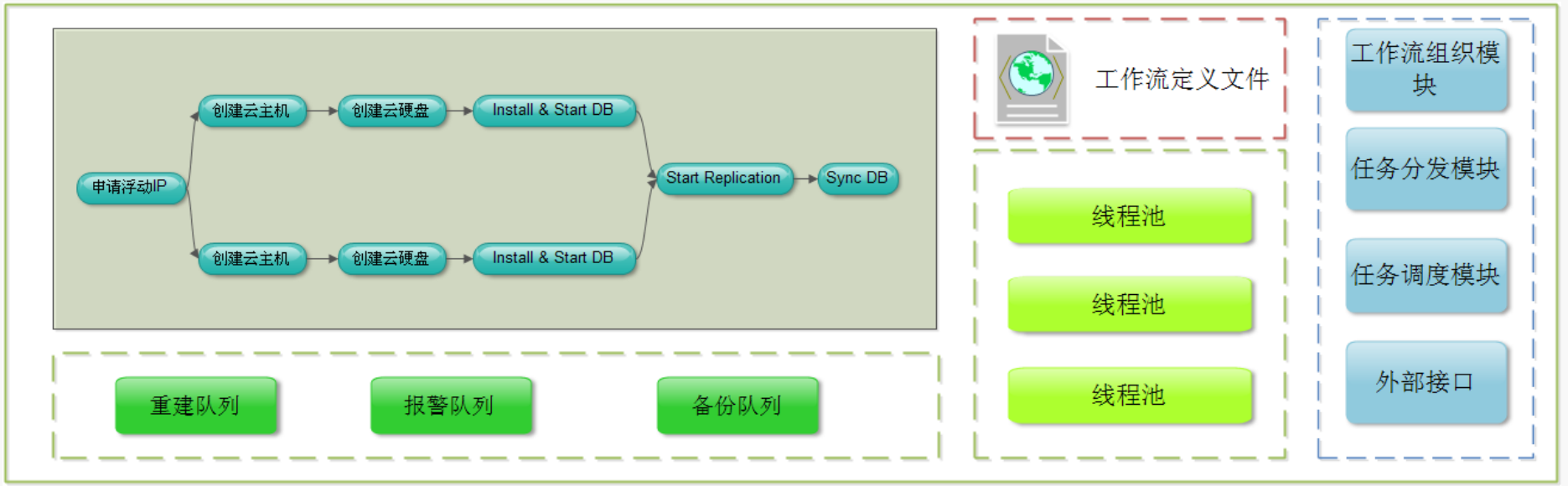
系统架构



管理服务器

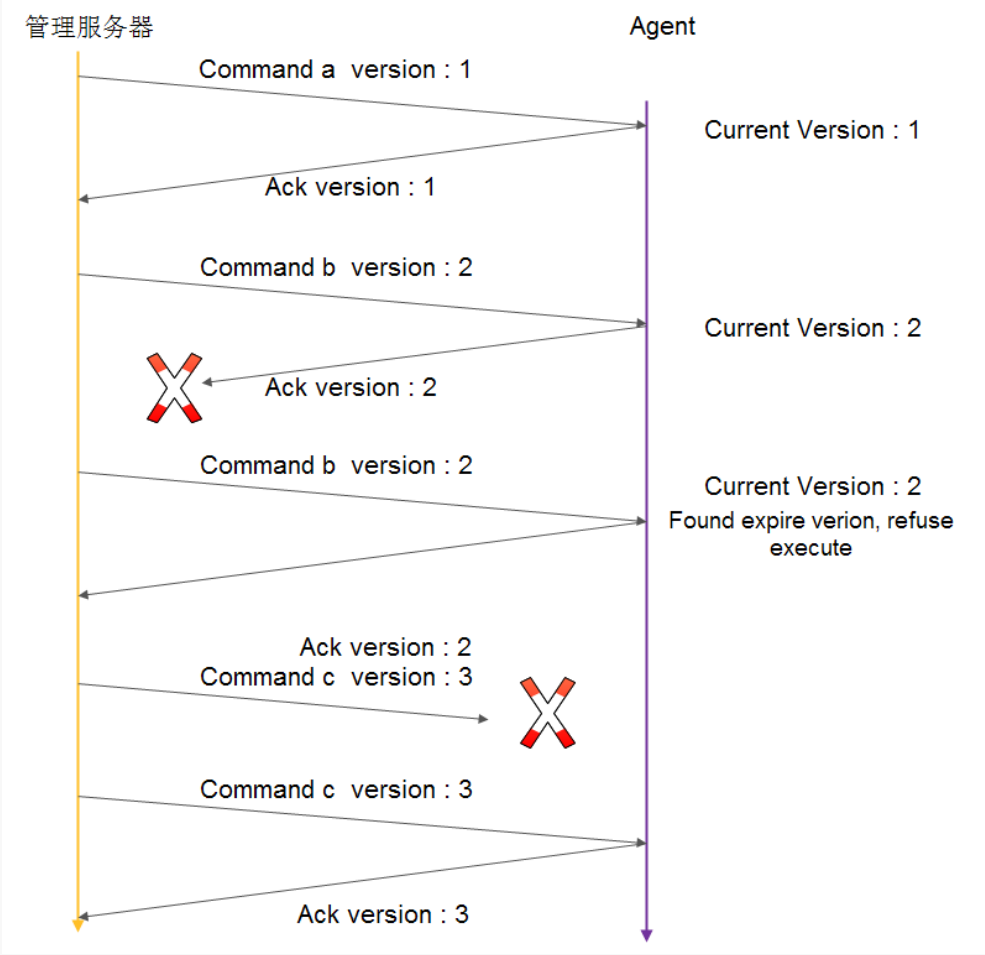
- 高可用
- 工作流
- 任务队列





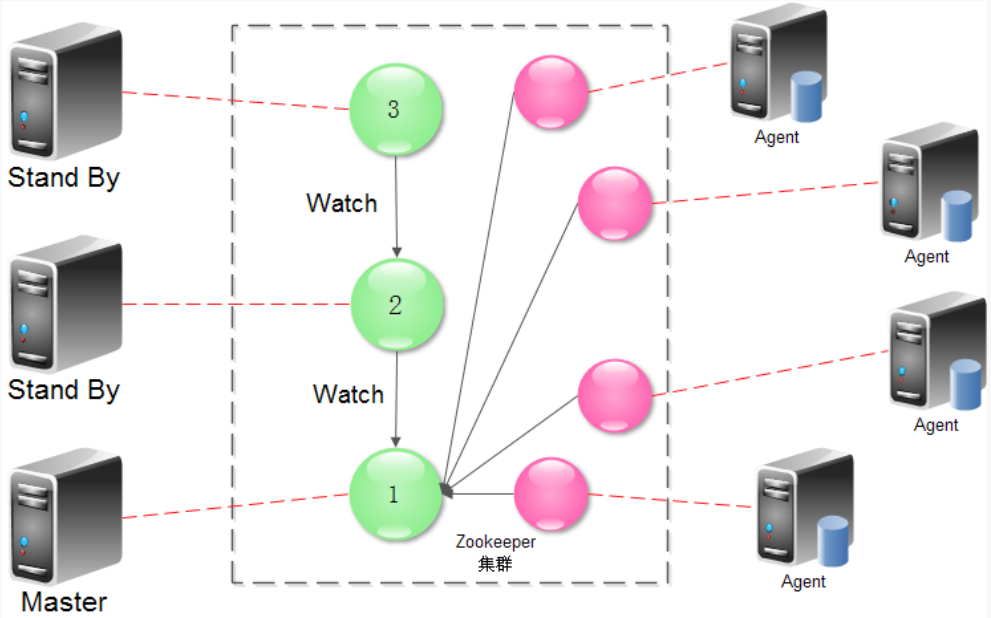
其他开源组件

- **ZeroMQ**
  - 轻量级
  - 异步传输
  - 消息序列化
  - 消息持久化



其他开源组件

- **Zookeeper**
  - 选主
  - 分布式锁
  - 配置管理



目录

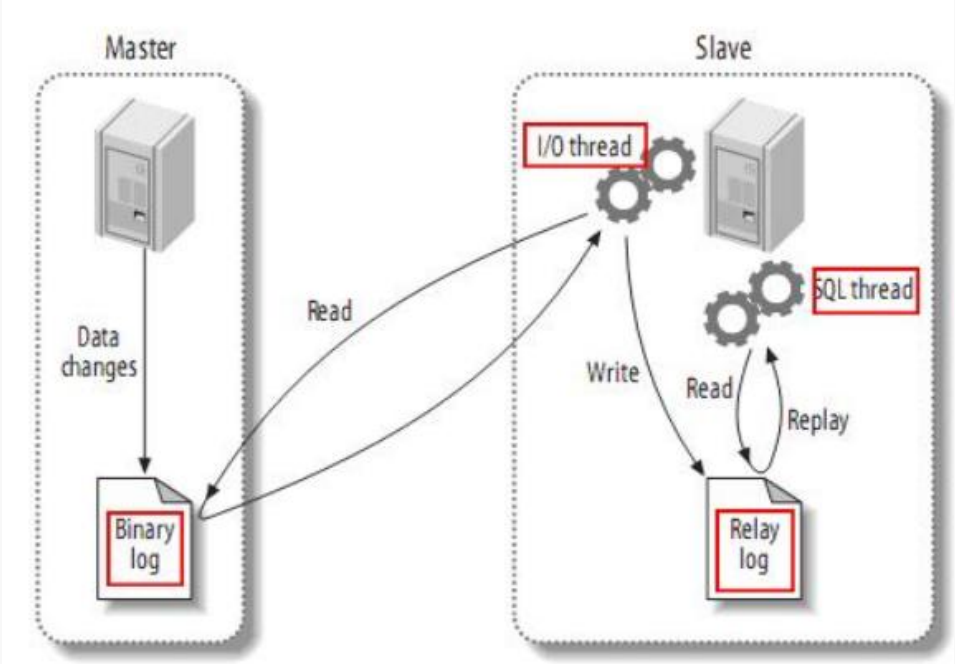
- 网易云数据库
- 架构架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

现有业界高可用方案

	可用性（停服时间）	可靠性（数据一致性）	对性能损耗	对应用透明	其他限制
DRBD	低	高	有	是	配置管理复杂
共享磁盘	低	高	无	是	无
binlog高可靠	高	高	有	是	需额外开发binlog监听工具
MMM	高	低	无	是	无
MHA	高	高	无	是	主机binlog是高可靠的

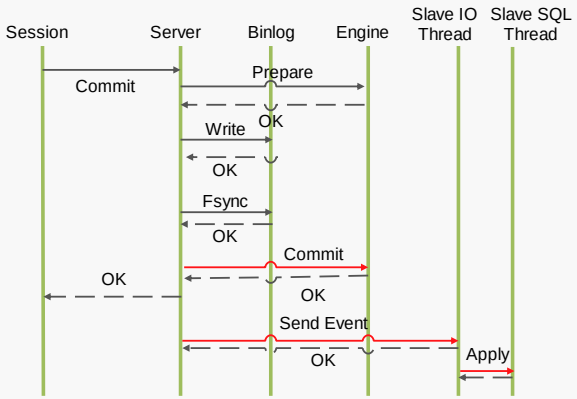
MySQL复制

- **主要挑战：**
  - **数据一致性**
  - 实时切换

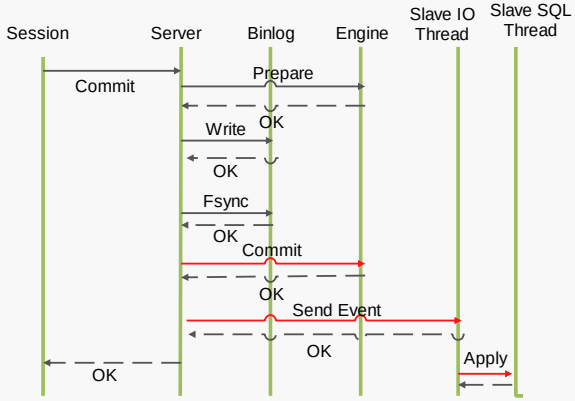


数据一致性

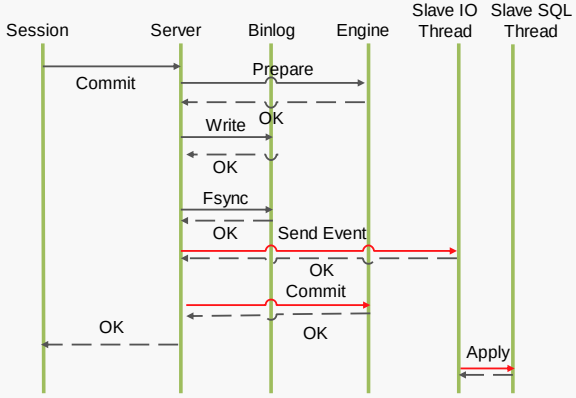
- 虚拟同步复制



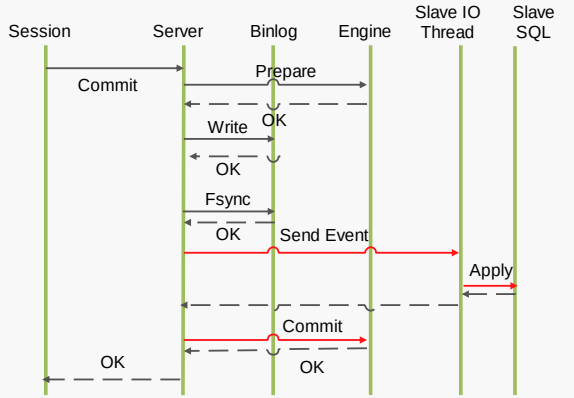
异步复制



半同步复制



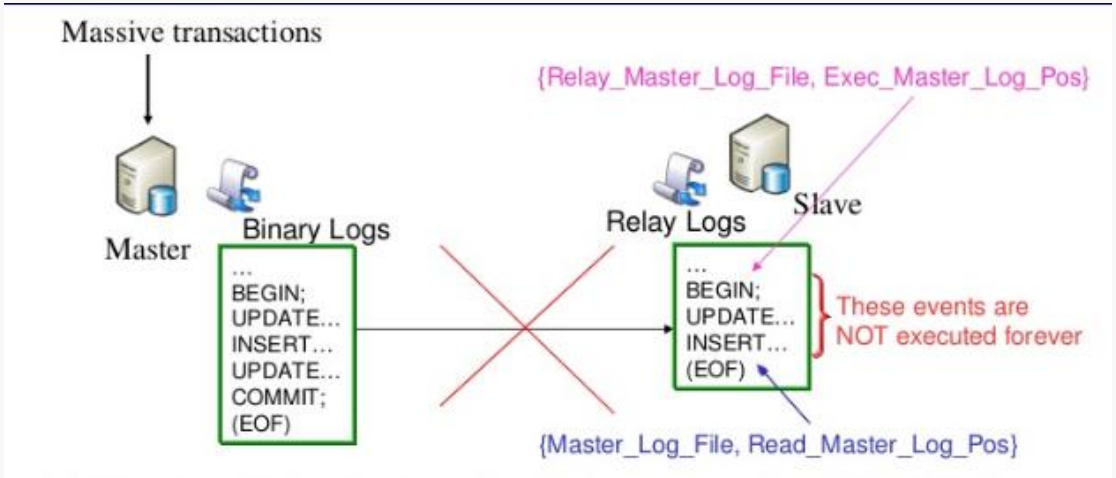
虚拟同步复制



全同步复制

数据一致性

- **IO Thread 隐患**
  - relay_log_recovery
- **SQL Thread**
  - Crash safe
- **Partial Transaction**

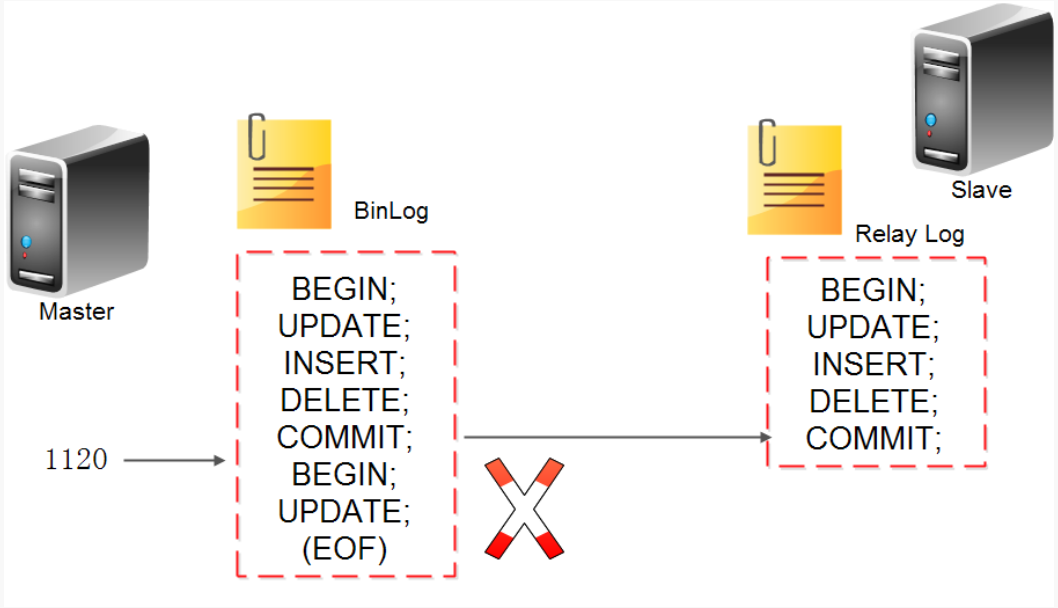


实时切换

- Batch Commit
- 并行复制
  - MariaDB
  - MySQL 5.7

修复

- binlog裁剪



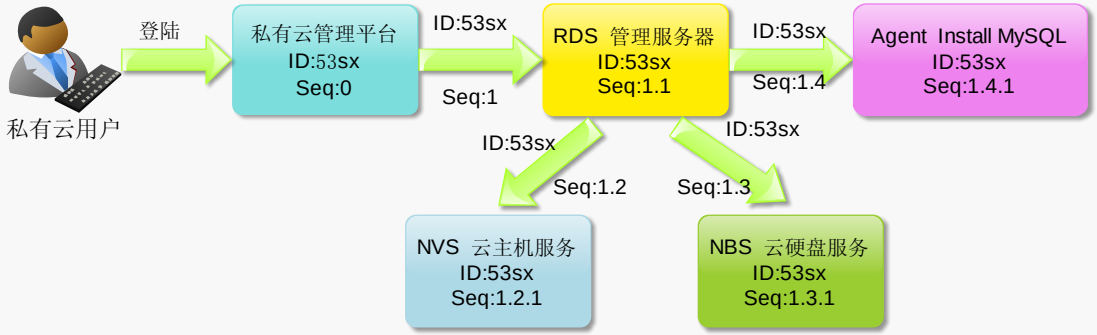
目录

- 网易云数据库
- 系统架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

系统监控

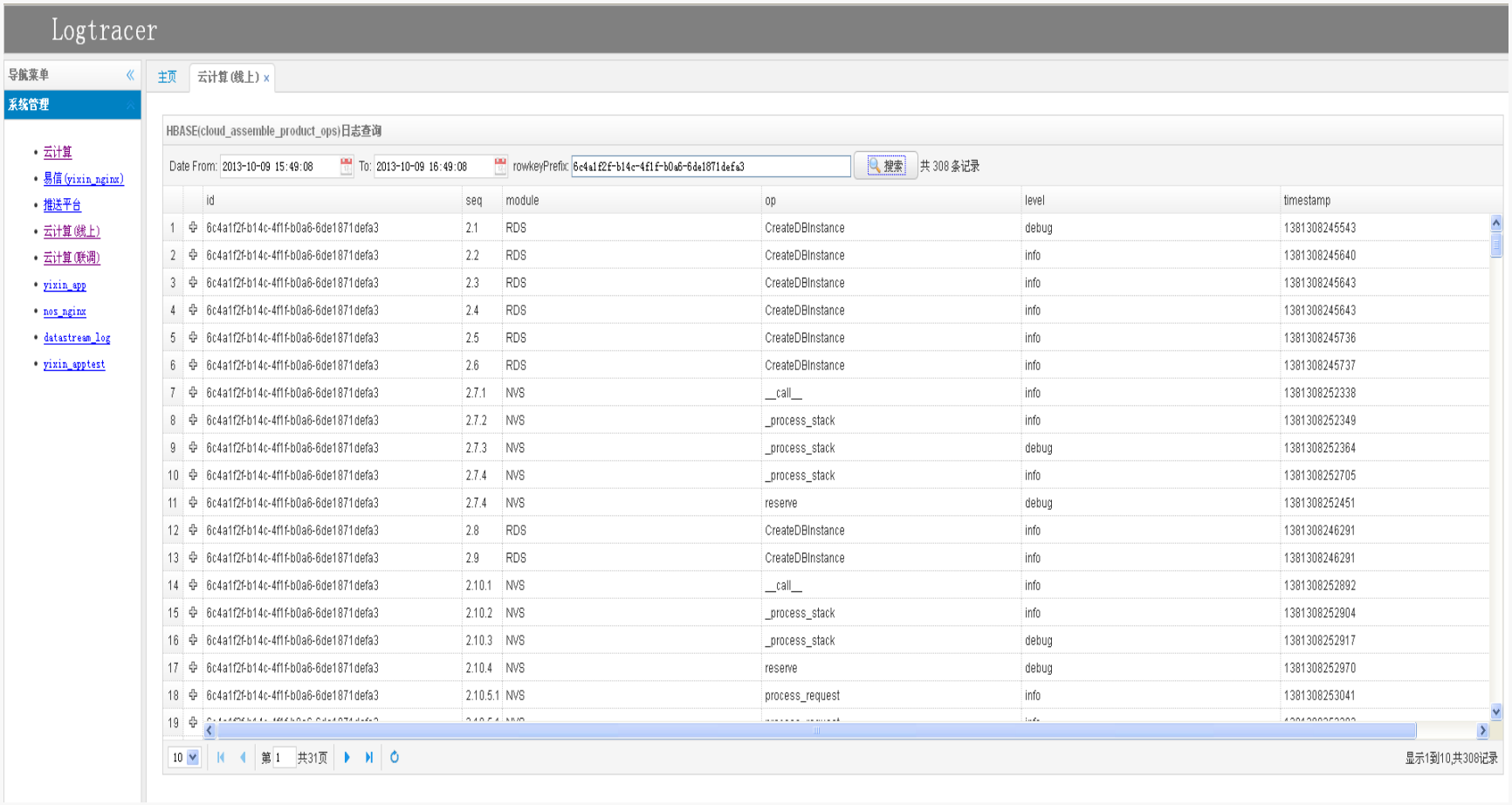
- **监控方式**
  - **主动**
  - **被动**
- **统一日志**

Key	Value
ID	用户请求唯一标识符
Seq	调用序号，格式为“x.x.x”，同一请求，模块内部每次写入日志或者调用其他模块前加1，跨模块加层，在服务和模块之间传递
Timestamp	时间戳
Level	日志输出级别
Module	模块名称
IP	写入日志的服务器
Identifier	用户标识
Op	操作名称
Object	对象信息，Json格式
Description	详情，异常需包含全部堆栈信息



網易 网易云 beta cloud.netease.com

系统监控

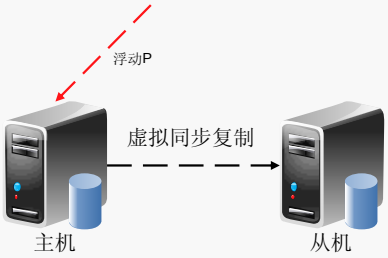


目录

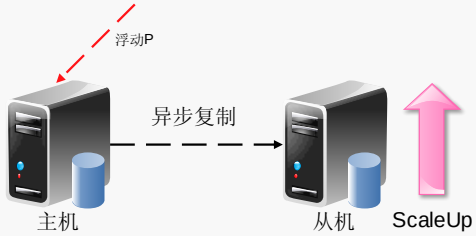
- 网易云数据库
- 架构架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

扩容

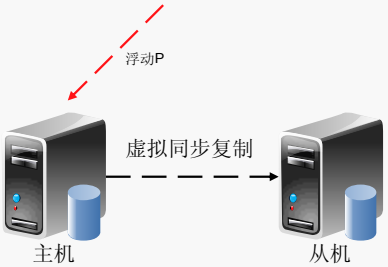
- 在线Scale Up
- 在线Scale Out



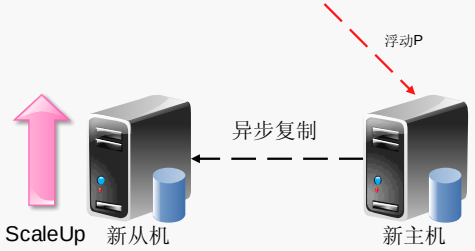
（1）主备虚拟同步复制



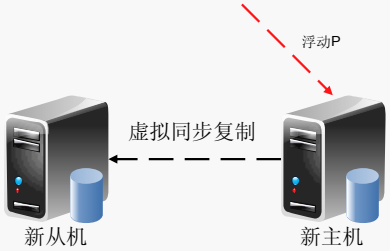
（2）主机切异步复制、从机ScaleUp



（3）重新建立虚拟全同步复制



（4）主从切换、新从机ScaleUp



（5）建立新从机到新主机的虚拟同步复制

目录

- 网易云数据库
- 架构架构
- 高可用设计
- 监控运维
- 在线ScaleUp和ScaleOut
- 未来与展望

- 未来工作
  - 多版本管理
  - LXC（Linux Container）
  - 支持Memcached