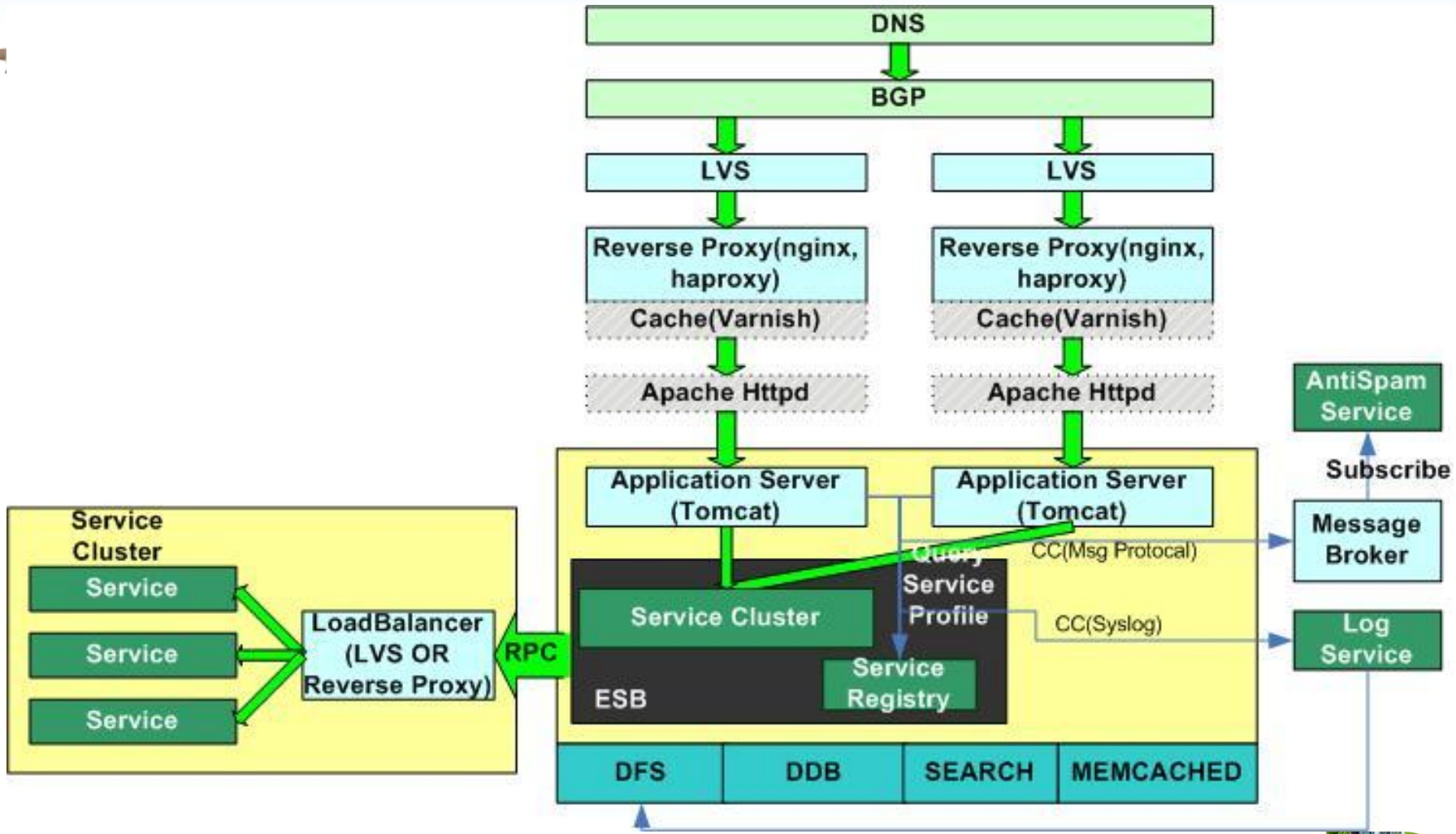
網易 NETEASE
www.163.com

网易大型应用架构与实践

汪源 网易.杭州研究院.副院长

应用架构与软件解决方案



硬件解决方案

- 大量使用廉价低端服务器
- 主要机型
  - 逻辑处理/缓存：刀片，如IBM HS22，24/48G内存
  - 数据库：2U机架式服务器，如Dell R710，单机多RAID 1多数据库进程
  - 文件/数据库备份：定制存储服务器（后将祥述）
- 其它
  - 网络：全部用千兆网络
  - SSD：邮件、搜索、数据库中少量采用SSD，如Intel 320
  - Amazon EC2：海外服务器

ESB: 目标与背景

- 目标
  - 解决解耦，独立部署
  - 基于服务的系统整合：SOA
  - 服务过载控制
- 解决的常见问题
  - 服务器位置变更后通知客户端
  - 同类服务多机提供负载均衡
  - 不同应用/平台/语言用不同服务调用方式
  - 多线程模型（用线程和线程池来处理服务调用）高并发性能问题
  - 服务过载导致系统失去响应

ESB: 设计

- 中心注册机制：服务提供方注册服务信息，服务需求方问中心注册机制获取提供方位置
- 通过Adapter/Stub机制实现，应用开发者只需要像调用普通Java方法一样调用Service API，针对不同协议的Adapter实现远程调用
  - 无论用什么协议调用代码相同
- 支持http/hessian/rmi/socket/jms/web service等调用协议
- 负载控制：控制请求队列长度及过载处理方式（抛弃老请求、拒绝新请求等）

DDB: 概述

- 目标
  - 保持关系数据库强大处理能力（如JOIN、批处理、事务ACID）
  - 提供一定的可伸缩性（百台）和可用性（无整体SPOF）
- ROADMAP
  - 2005年底启动，2006.9.1随博客正式上线
  - 5年持续开发，最新版本4.4
    - 近期新功能：支持多语言、主从自动切换、集成Memcached
- 应用情况
  - 为公司几乎所有WEB类应用采用：邮箱、博客、微博、POPO等
  - 总体规模：数据库节点400+，数据量60TB

DDB: 功能与特色

- 基于“实体组分区映射”实现的可伸缩与负载均衡
- 多平台和多语言
  - Java提供定制JDBC驱动，其它语言通过MySQL协议
- 支持常用的RDBMS功能
  - 如支持JOIN、分组、排序、聚焦函数、视图、存储过程、触发器、用户权限
- 跨节点或跨DDB分布式事务(2PC)
- 在线模式修改
- 支持MySQL和Oracle混合使用
- 丰富的管理功能，GUI/CLI/WEB管理工具
  - 模式定义、权限管理、统计、数据迁移、定时任务等等

DDB: 分区与可伸缩性

- 分区策略
  - 表按某指定属性的值进行分区：哈希、自定义函数
  - 分区映射：分区到节点的映射表缓存于所有DDB Driver，高效的直接路由。调整映射表实现扩容和负载均衡
  - 实体组：多表可参与同一分区策略，减少分布式JOIN和分布式事务
- 系统扩容
  - 借助MySQL复制与DDB查询改写实现，在线成倍与非成倍扩容
  - 成倍扩容
    - 建立待扩容节点A的两个镜像A1和A2，在A1和A2上删除多余的分区数据
    - 等A1和A2与A接近同步后短期停止对A的访问，等A1和A2与A完全同步后，修改分区映射（通知所有客户端）后A下线，A1与A2上线
    - DDB Driver转入迁移中处理模式，为有关SQL加上分区选择条件，防止触及A1/A2上的多余分区数据
    - A1/A2再次清理少量多余分区数据后DDB Driver转入正常处理模式
  - 非成倍扩容
    - 轮流复制，以2节点扩容到3节点为例，新增节点轮流与原两个节点复制，同样短期停止访问后修改分区映射

DDB: 分布式事务

- 使用两阶段提交
- 内嵌于应用的DDB JDBC Driver为协调者TM，后端数据库为RM
- TM崩溃导致分布式事务处于悬挂状态，导致记录长时间被锁定？
  - TM写分布式事务LOG，重启根据LOG处理
  - MASTER定期检查回滚悬挂事务
- MySQL bugs
  - 两阶段提交实现不符合规范：修改MySQL代码，在连接断开时不回滚已经处于PREPARED状态的事务
  - XA事务出现错误后继续执行SQL导致MySQL崩溃：DDB Driver层禁止出现错误后继续执行SQL（最新MySQL已修复此问题）

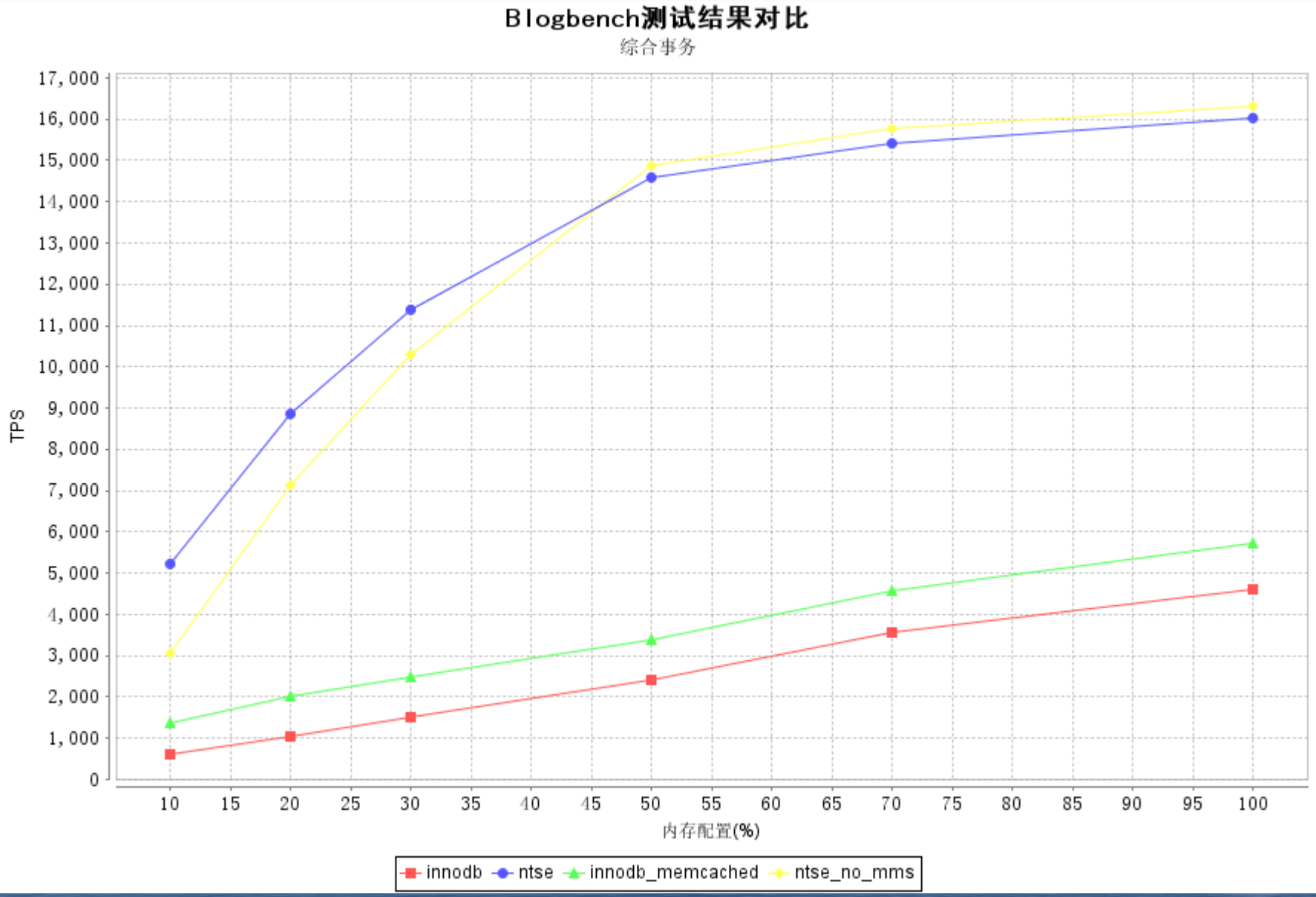
数据库集成Memcached

- 方案一：DAO框架
  - 框架提供CRUD类辅助接口，应用基于这些接口方便实现有无Memcached两套DAO
  - 通过Spring框架动态配置选择用哪套DAO实现，细化到具体的DAO类
- 方案二：DDB集成Memcached
  - 通过DDB配置各表是否启动Memcached缓存功能
  - DDB Driver根据SQL语义智能化同步操作数据库与Memcached
- UPDATE时都从Memcached中删除，减少Memcached与数据库不一致概率

NTSE

- 面向WEB应用特征开发的高性能MySQL存储引擎
- 支持单记录操作ACID，不支持多操作组合事务
- 功能特色
  - 行级缓存有效缓存热点记录
  - 索引前缀压缩：索引大小约InnoDB 30%
  - 大对象LZO压缩：35%-40%
  - 基于字典的记录压缩：50%
  - 在线建索引
  - 高性能KV风格API：如SQL操作同一份底层数据，支持binlog
- 应用情况：应用于新闻跟帖、反垃圾、轻博等系统，近20个节点，1T数据

NTSE vs InnoDB: 10X



TNT

- 基于NTSE，通过多版本支持ACID事务
- 内存多版本，外存单一版本（几乎原封不动的NTSE）
  - 没有类似于InnoDB或PostgreSQL的版本化信息开销
- UPDATE/DELETE缓存，有助于变随机IO为顺序IO
- 开发中。。。

NoSQL&NewSQL

- 另一套类似于Google的Infrastructure：有道
  - ODFS→GFS、OMAP→BigTable、CoWork→MapReduce
- 对NoSQL和NewSQL的需求不强烈
  - 网易的数据库应用规模较小：最大50节点左右
  - DDB基本能满足需求：在线扩容、在线改模式
  - NoSQL不一定性能高
    - 数据量大于内存时通常性能不高，如TC、MongoDB
- 少量使用
  - TokuDB：超快INSERT，但UPDATE/DELETE/SEARCH性能都差
  - MongoDB：用户Profile等树状数据
  - Redis：更好的Cache

DFS

- 目标：中型（几K~10MB）文件存储，主要用于相册和网盘
- 提供基于文件ID的CRUD接口
- 功能特色
  - 分区与可伸缩：实现与DDB类似
  - 可配置任意数量复本，2PC保证多复本一致
  - MD5去重：20-40%重复
  - 多DC部署
  - 存储节点HTTP下载（含权限检查）
- 应用情况
  - 应用：网易相册、网盘、超大附件等
  - 规模：节点数1200+，数据量4PB+

DIR

- 基于Java Lucene、Solr、Zoie
  - Solr：提供RESTful API
  - Zoie：实时全文检索
- 功能特色
  - 通过ETL平台与MySQL数据库建立对应关系，应用无需更新全文检索
    - 自解析binlog实现
  - 分区与可伸缩：实现类似于DDB的分区映射
  - 同时支持全站和个人数据检索：全站搜索所有数据分区，个人根据分区策略定位一个节点

通用事件推送系统

- Timeline定义：所follow的用户的事件按某种规则排序后的事件列表
- 用途：所有类“好友动态”类应用，如博客、轻博、微博、SNS等
- 灵活通用：多种timeline，多种排序规则
- 性能优化
  - Push/Pull相结合，明星Pull，普通用户Push
    - 纯Pull：follow大量用户时Pull合并性能问题
    - 纯Push：明星有大量粉丝时性能问题
  - 事件列表分组压缩存储、利用NTSE更新缓存优化
  - 优先推送给在线用户

定制存储服务器

- 核心：使用桌面级硬盘，节约成本
  - 节约成本约30%
- 国内硬件产商联合研发定制，2U 12盘或4U 24盘
- 轮休机制
  - 问题：桌面级硬盘7*24工作导致故障率增加
  - 方案：
    - 数据更新先缓存于企业级硬盘存储集群，定期合并
    - 多复本：一般部署2复本
    - DFS软件支持，定期禁止访问某一复本
    - 硬件支持，无IO时自动休眠
- 应用于邮箱、相册、网盘等，规模30PB+