淘宝网架构变迁和挑战

TAOBAOJM

About me

- 姓名:曾宪杰
- 花名:华黎
- 淘宝-产品技术-Java中间件团队
- 团队博客 http://rdc.taobao.com/team/jm/
- Sina微博 @曾宪杰_华黎
- Twitter @vanadies10

内容提要

- 淘宝网架构的变迁
- 通用的基础产品
- 我们面临的挑战

淘宝网的架构变迁

TAOBAOJM

技术里程总览

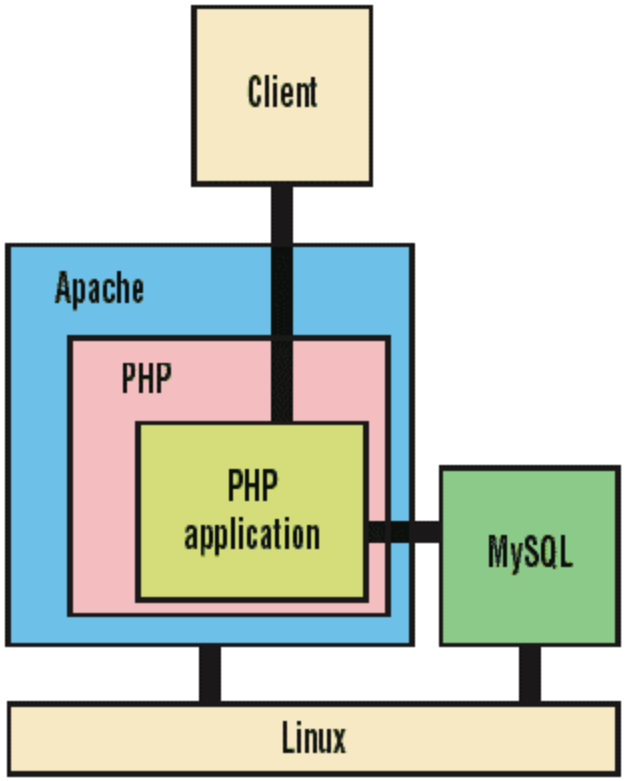
- V1.0 （2003.5 – 2004.5）
  - ✓ 小而快的简单架构
- V2.0 （2004.2 – 2008.3）
  - ✓ 应付业务增长的层次化结构
  - ✓ 开始有自己开发的软件技术
- V3 （2007.10-2009.11）
  - ✓ 产品化的思维
    - ✓ 做深，做精
  - ✓ 服务导向的框架
    - ✓ 应付多样化的业务
- V4 （2009.8 - ）
  - ✓ 系统化
  - ✓ 智能化
  - ✓ 专业化

V1.0：小而快
2003.5 – 2004.5

TAOBAOJM

简介

- 2003年非典时期
- 使用LAMP架构（Linux, Apache, MySql, Php)
  - ✓ 业界流行的免费开源组合
- 使用改造过的一个商业软件
  - ✓ phpAuction，以拍卖为主
- 简单的库表结构
  - ✓ 用户，交易，列表，其他



简单的结构，但符合当时需求

V2.0：多层次结构，开始做自己的软件
2004.2 – 2008.3

TAOBAOJM

需求

- 业务发展快速
  - ✓ 需要“较高”性能的架构（百万至千万用户级的架构）
- 团队并行开发
  - ✓ 开始有小团队（几十人）做开发，开发效率必需考虑
- 系统的可伸缩
  - ✓ 容易的扩容，增加机器

问题 I

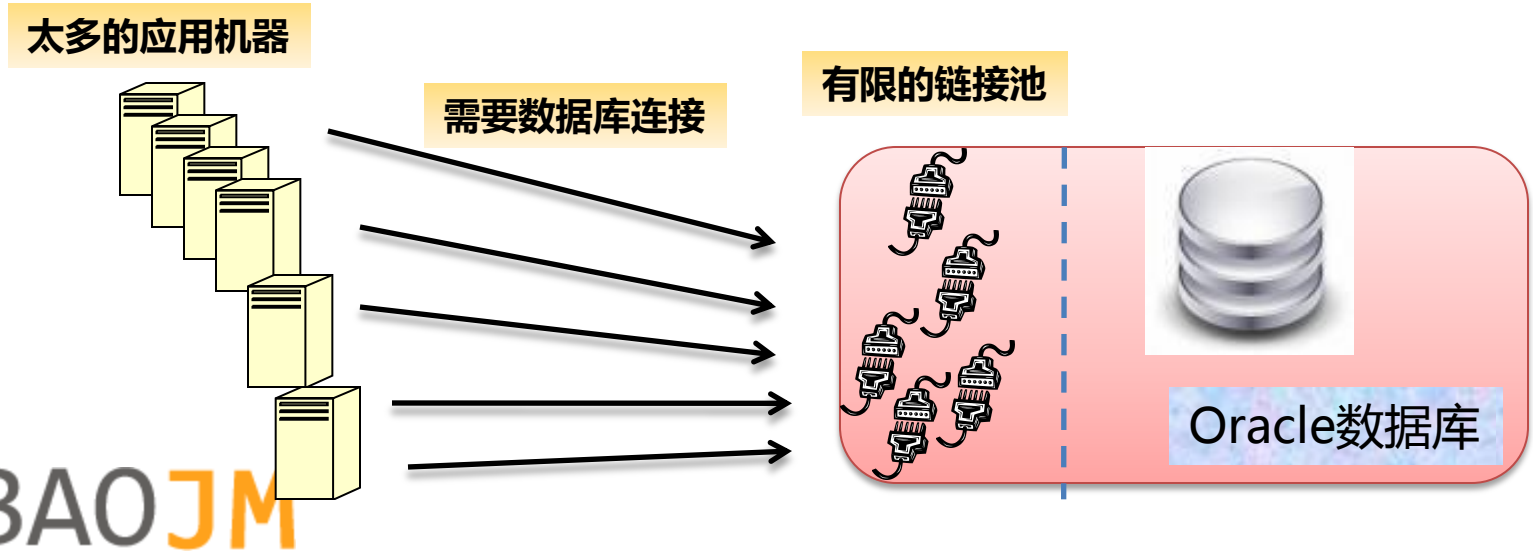
- 上百人维护一个代码百万行的核心工程
  - ✓ 共享一个代码模块，Denali（虽然部署是分离的）
- 多个业务系统中的超过1/3的核心代码重复编写

代码复杂难维护

V2单一代码模块的设计（Denali）

问题 II

- 所有系统都要关心数据拆分规则
  - ✓ 不必要的设计
- 数据库连接达到上限 （每个Oracle数据库大约提供5000个链接）
  - ✓ 连接池是有限的资源



V3.0：产品化思维及服务导向框架
2007.10-2009.11

TAOBAOJM

需求

- 支撑大型团队，丰富业务的并行开发
  - ✓ 软件模组化，中心化（用户，交易，商品，店铺，评价等），走向鬆耦合
  - ✓ 基础软件产品化
    - ✓ 独立团队开发，做深，做大
  - ✓ 从盖独立别墅到建造高楼大厦！
- 支撑高速的业务增长
  - ✓ 快速扩容（几十亿PV，几千亿GMV，几万台机器）
- 提高可用性及管理性
  - ✓ 走向Always Available
- 对外开放

V4.0：系统化、智能化、专业化
2009.8-

TAOBAOJM

现状

- 系统化
  - ✓ 把知识经验通过系统、平台进行沉淀，而不是总是人肉重复
- 智能化
  - ✓ 从提供开关人工处理到系统自主决策
- 专业化
  - ✓ 业务平台、技术平台的深耕
  - ✓ 稳定性、性能的深入发展

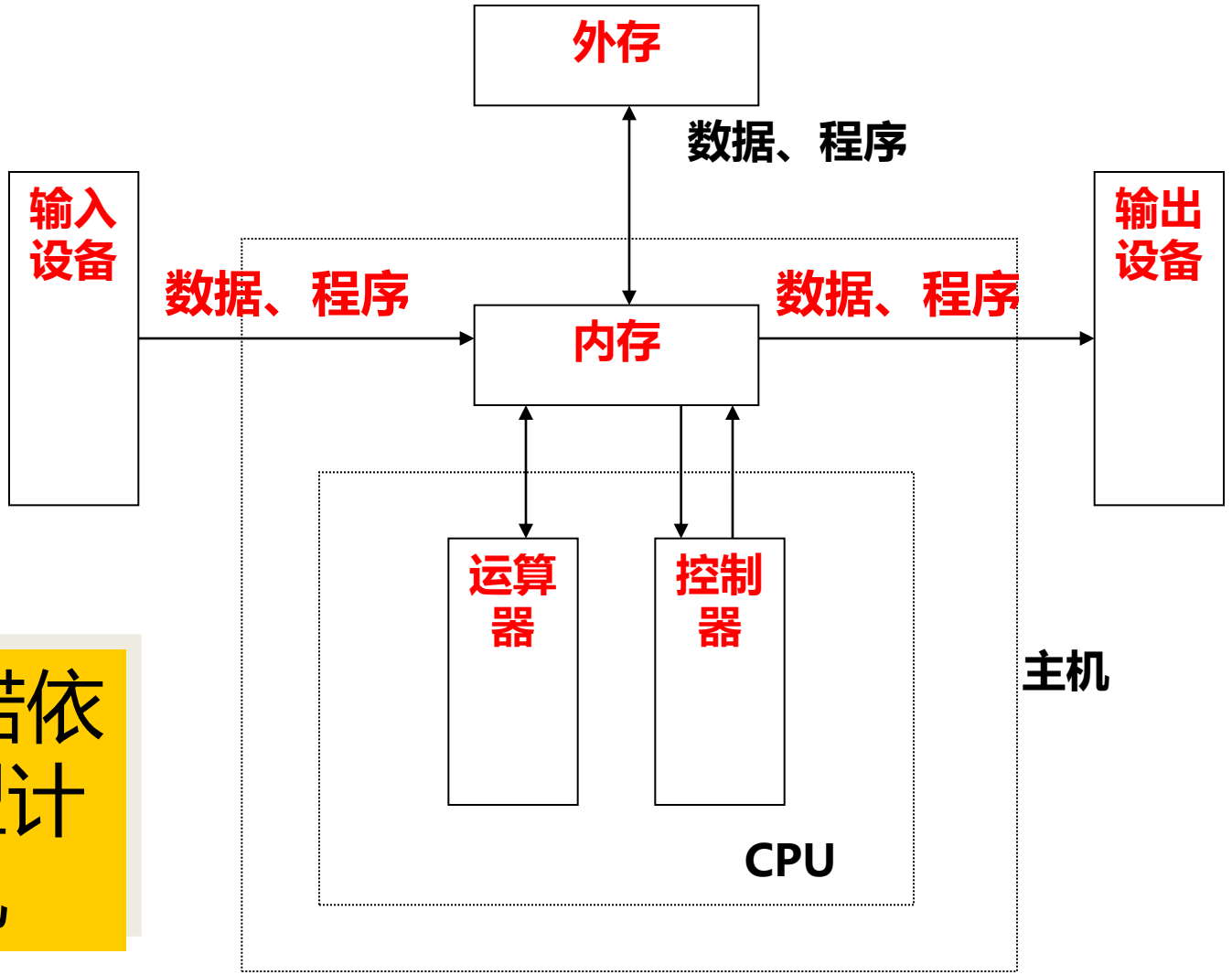
技术里程回顾

- V1.0 （2003.5 – 2004.5）
  - ✓ 小而快的简单架构
- V2.0 （2004.2 – 2008.3）
  - ✓ 应付业务增长的层次化结构
  - ✓ 开始有自己开发的软件技术
- V3 （2007.10-2009.11）
  - ✓ 产品化的思维
    - ✓ 做深，做精
  - ✓ 服务导向的框架
    - ✓ 应付多样化的业务
- V4 （2009.8 - ）
  - ✓ 系统化
  - ✓ 智能化
  - ✓ 专业化

通用的基础产品

TAOBAOJM

计算机组成



计算机硬件结构图

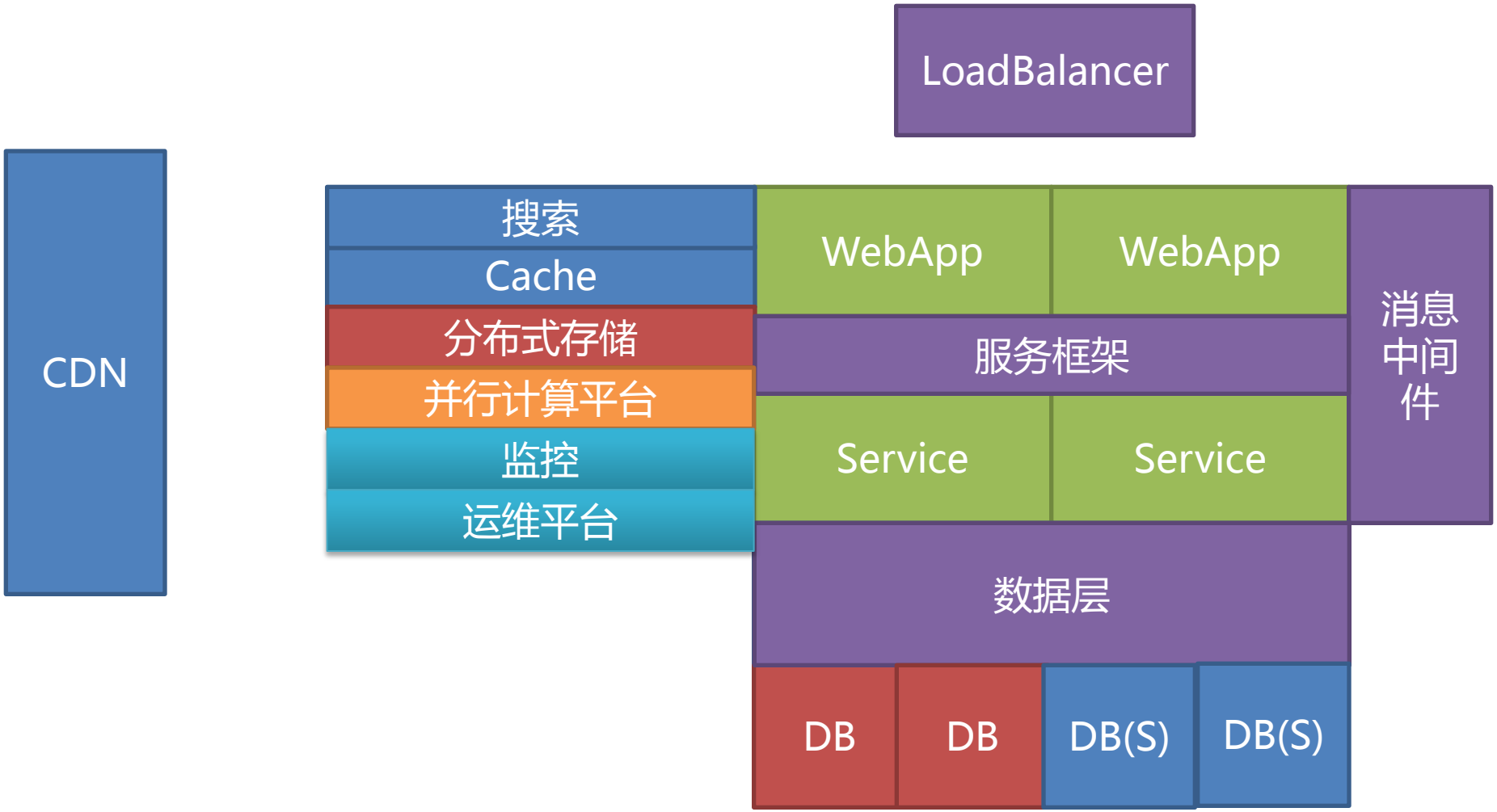
冯.诺依曼型计算机

计算机系统的本质

- 数据处理
- 数据存储
- 数据访问

网站结构示意图


SALE

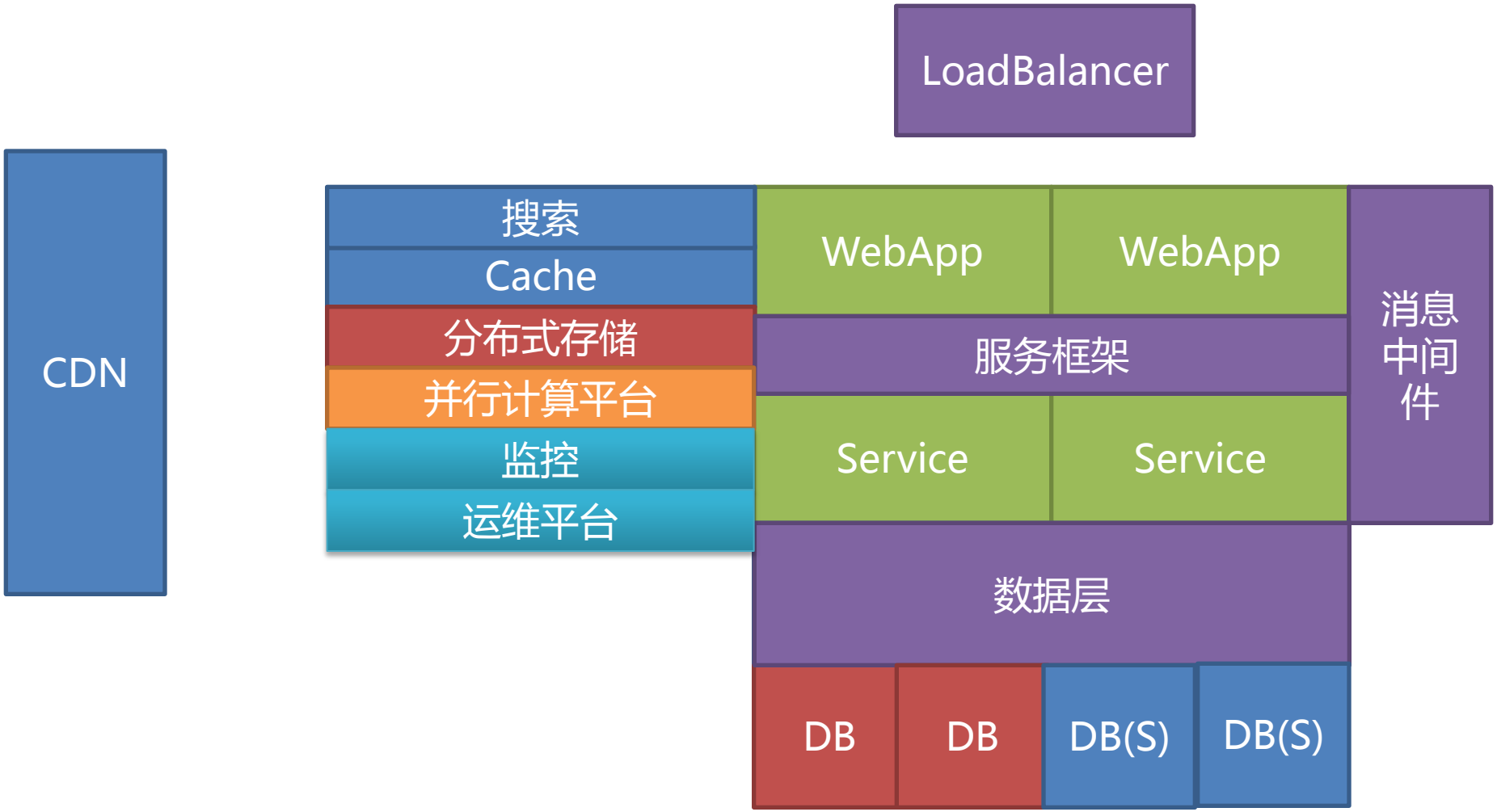

LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

Web框架

- Webx
  - ✓ 在阿里内部广泛使用的MVC框架
  - ✓ Turbine风格
  - ✓ 有很好的层次化、模块化，并且高度可扩展
- 基于Webx的无线应用自适应框架
- Velocity的编译优化

中间件


SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

服务框架

- 服务发布
- 服务查询
- 服务调用
- 服务治理

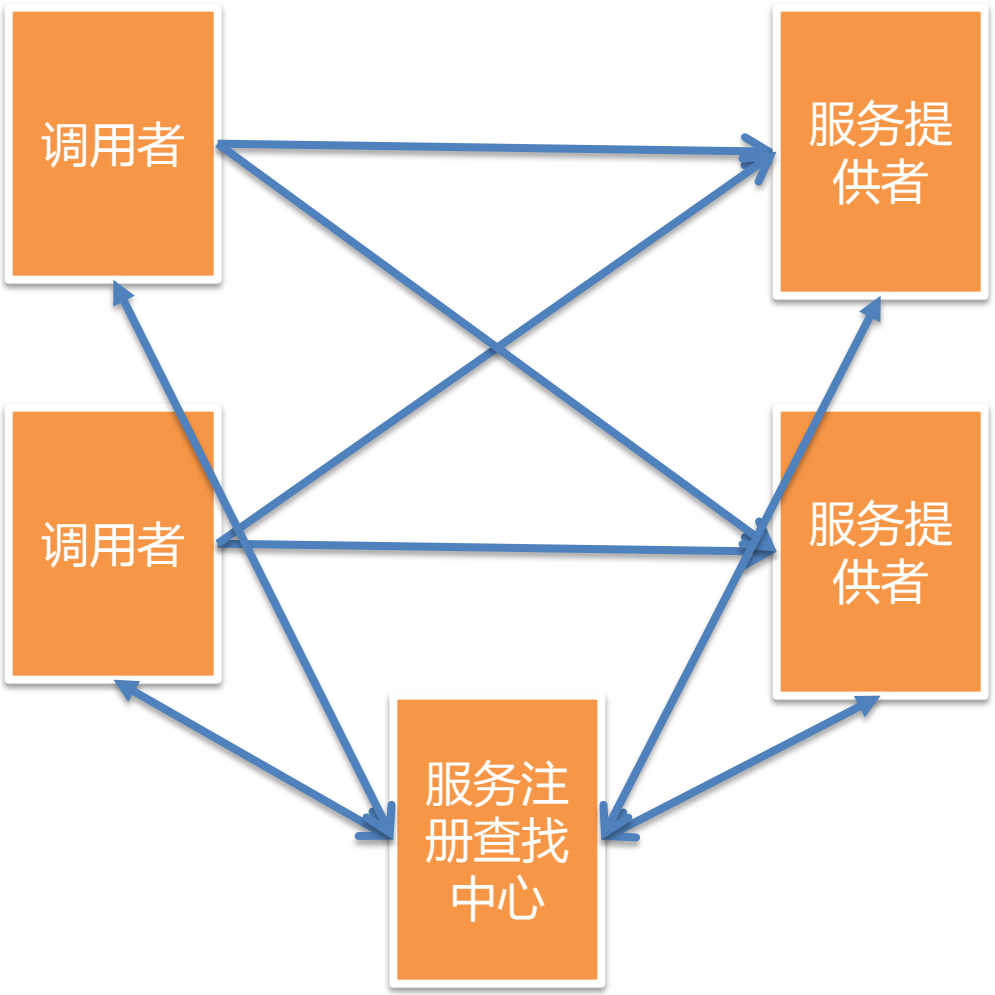
服务框架

- 07年淘宝开始走向服务化
- 08年初有多种RPC的方式
- 基于SOCKET，有多种不同实现
- 使用上不透明，成本高

服务框架

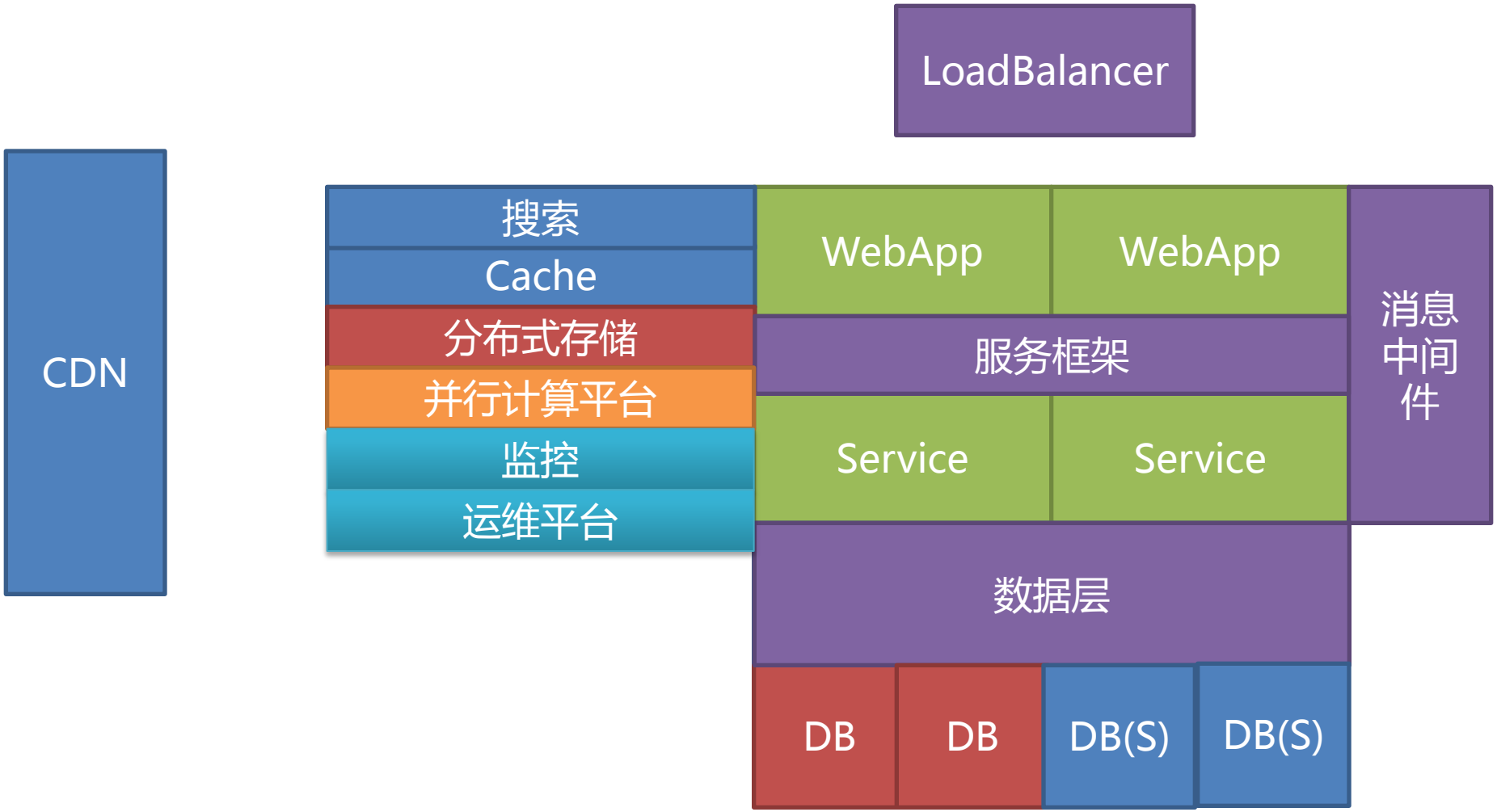
- 简单透明(提供服务和使用服务)
- 支持软负载
- 灵活可控，方便扩展
- 稳定性支持

服务框架



中间件

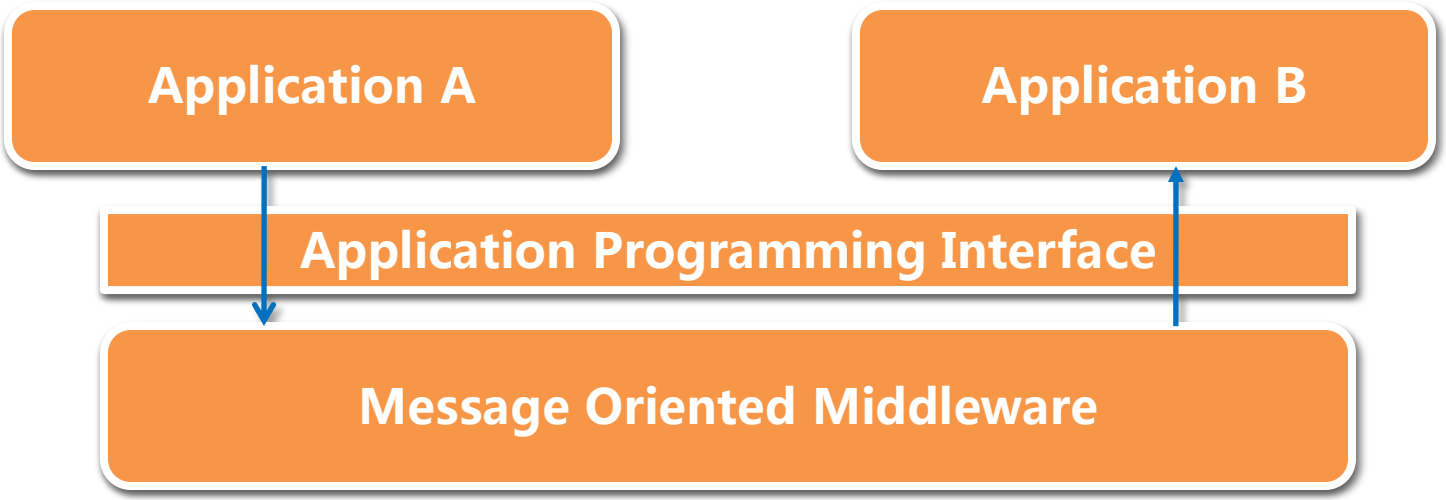

SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

消息中间件

- **Message-oriented middleware** (**MOM**) is software infrastructure focused on sending and receiving messages between distributed systems.

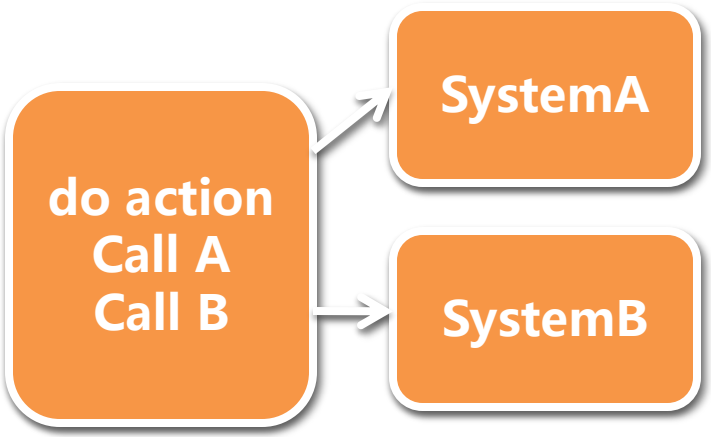
  --- from wikipedia.org



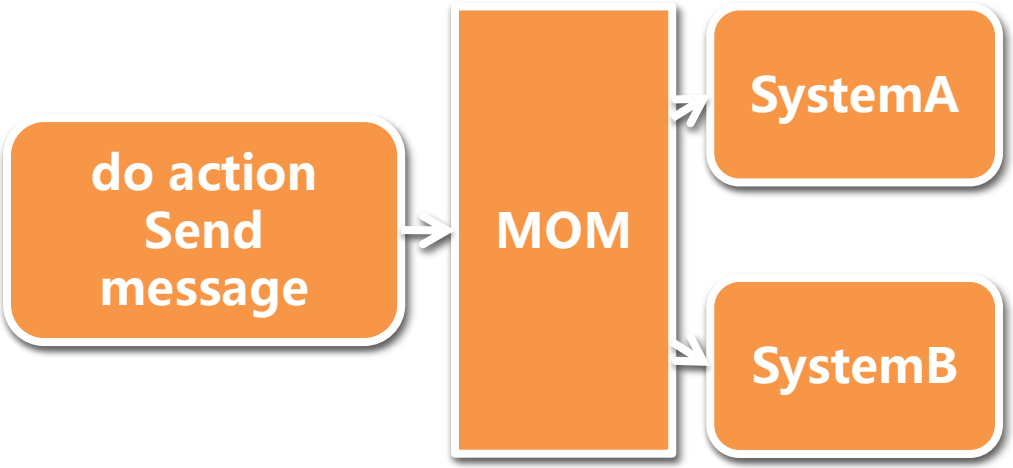
- MOM的优点
  - 松耦合
  - 异步处理

消息中间件

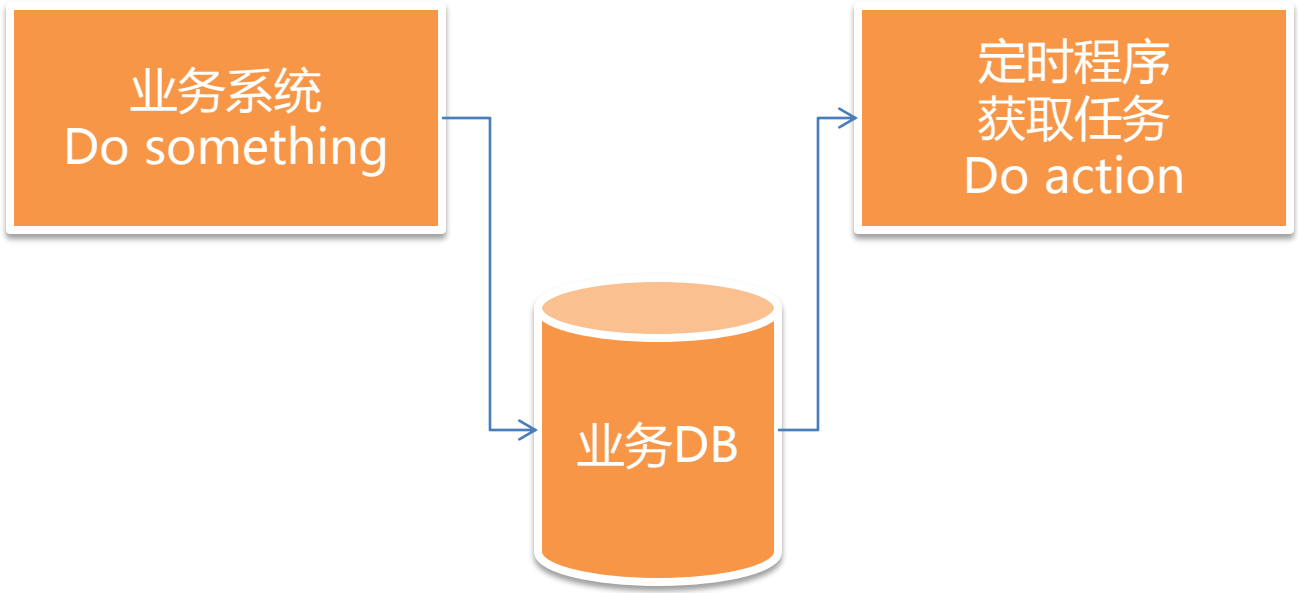
传统方式



使用消息中间件方式



消息中间件



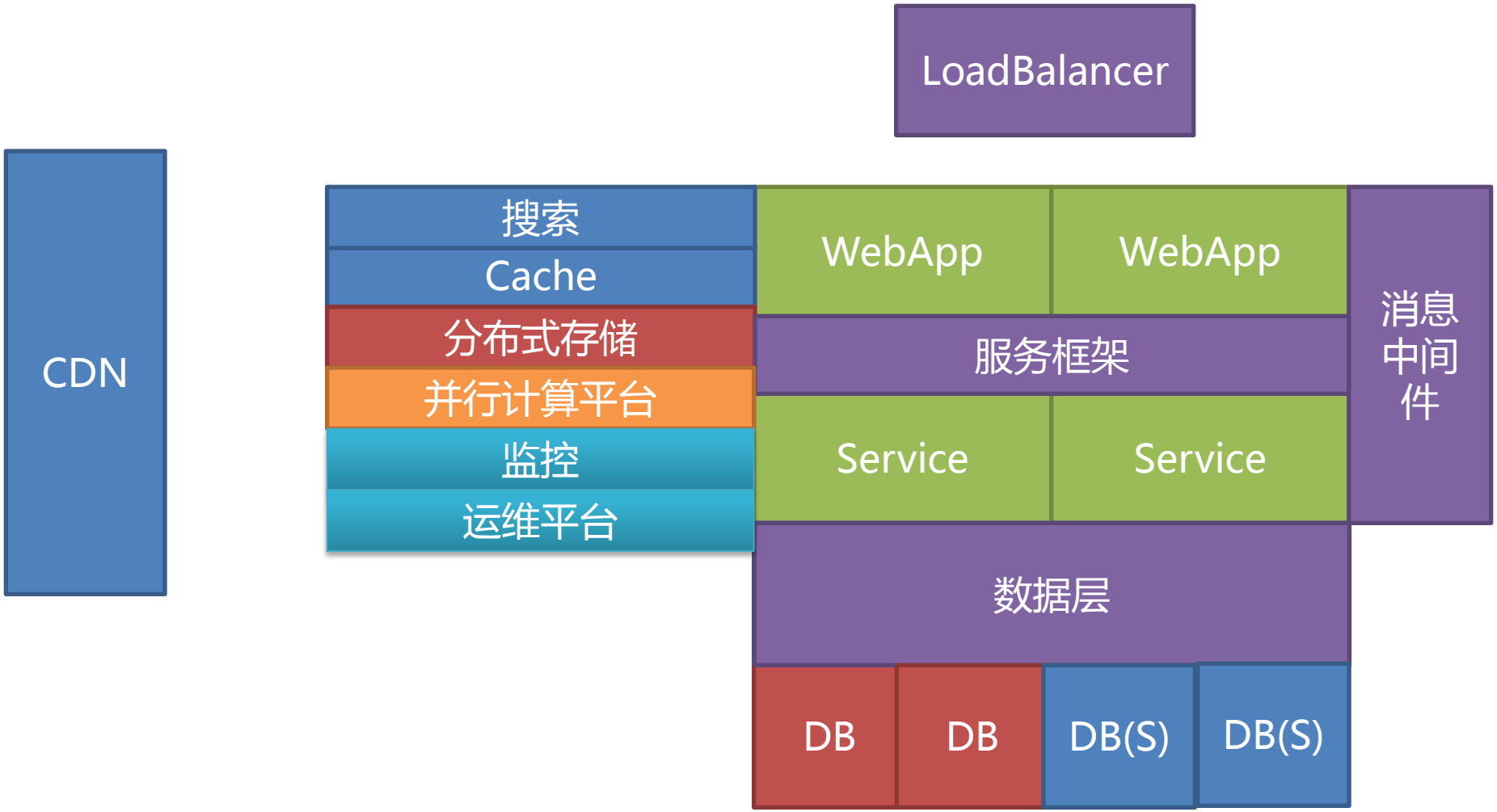
业务系统完成一件事情后，需要其他系统进行处理的，通过定时程序来驱动

消息中间件

- Notify是一个高性能、可靠、可扩展、可与发送端业务逻辑相结合、支持订阅者集群的消息中间件。
- 互联网时代的消息中间件
- 支持最终一致
- 消息可靠
- 支持订阅者集群

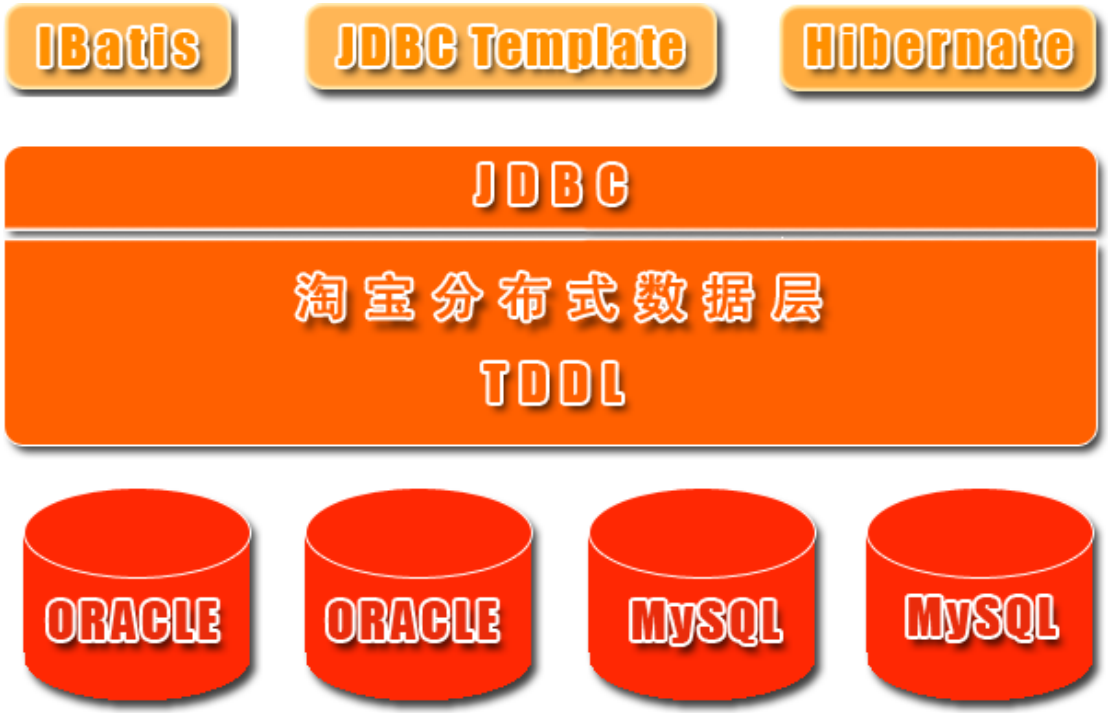
中间件


SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

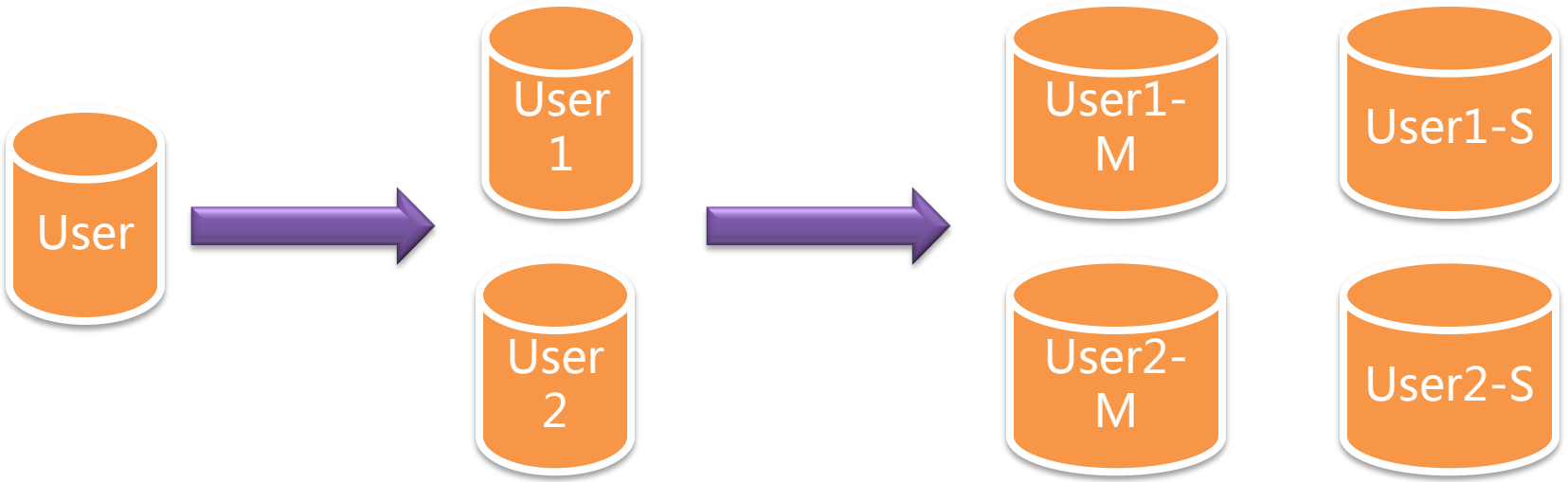
数据层


SALE

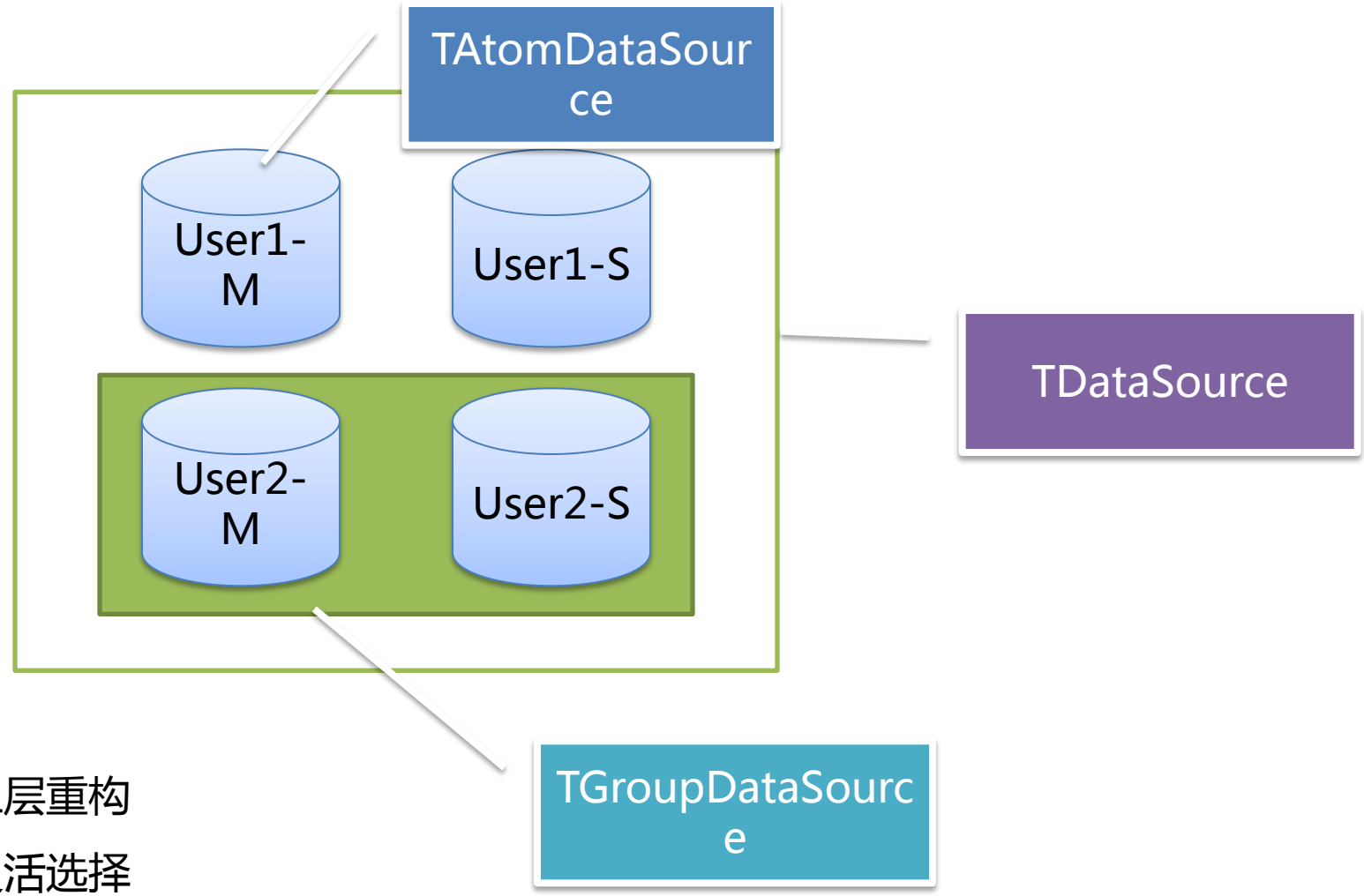

IBatis
JDBC Template
Hibernate
JDBC
淘宝分布式数据层
TDDL
ORACLE
ORACLE
MySQL
MySQL

数据层

数据库架构的演进



数据层



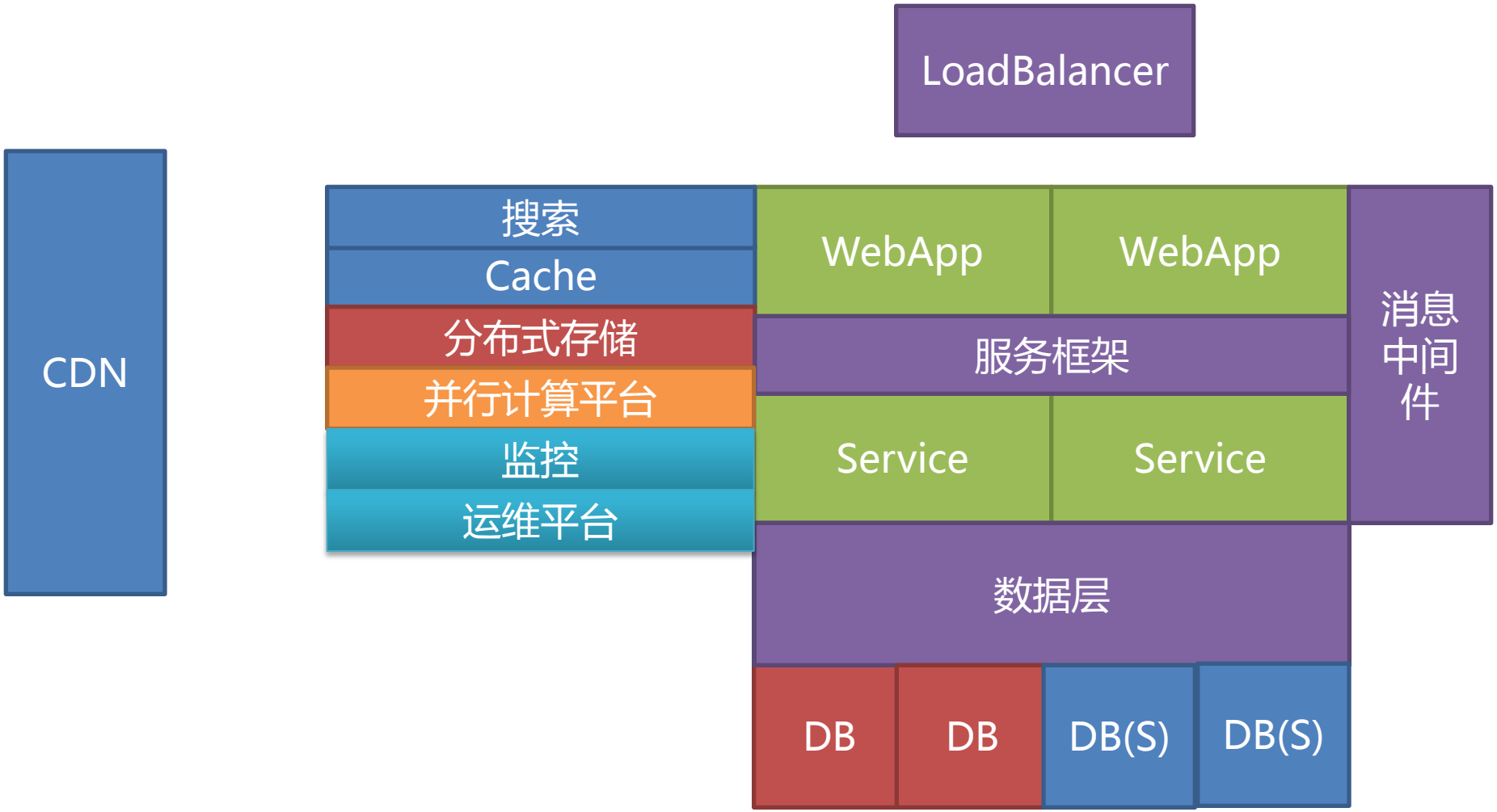
数据源的三层重构

业务可以灵活选择

数据层

- SQL解析，路由规则，数据合并
- Client->DB和Client->Server->DB模式
- 非对称数据复制
- 三层的数据源结构


SALE

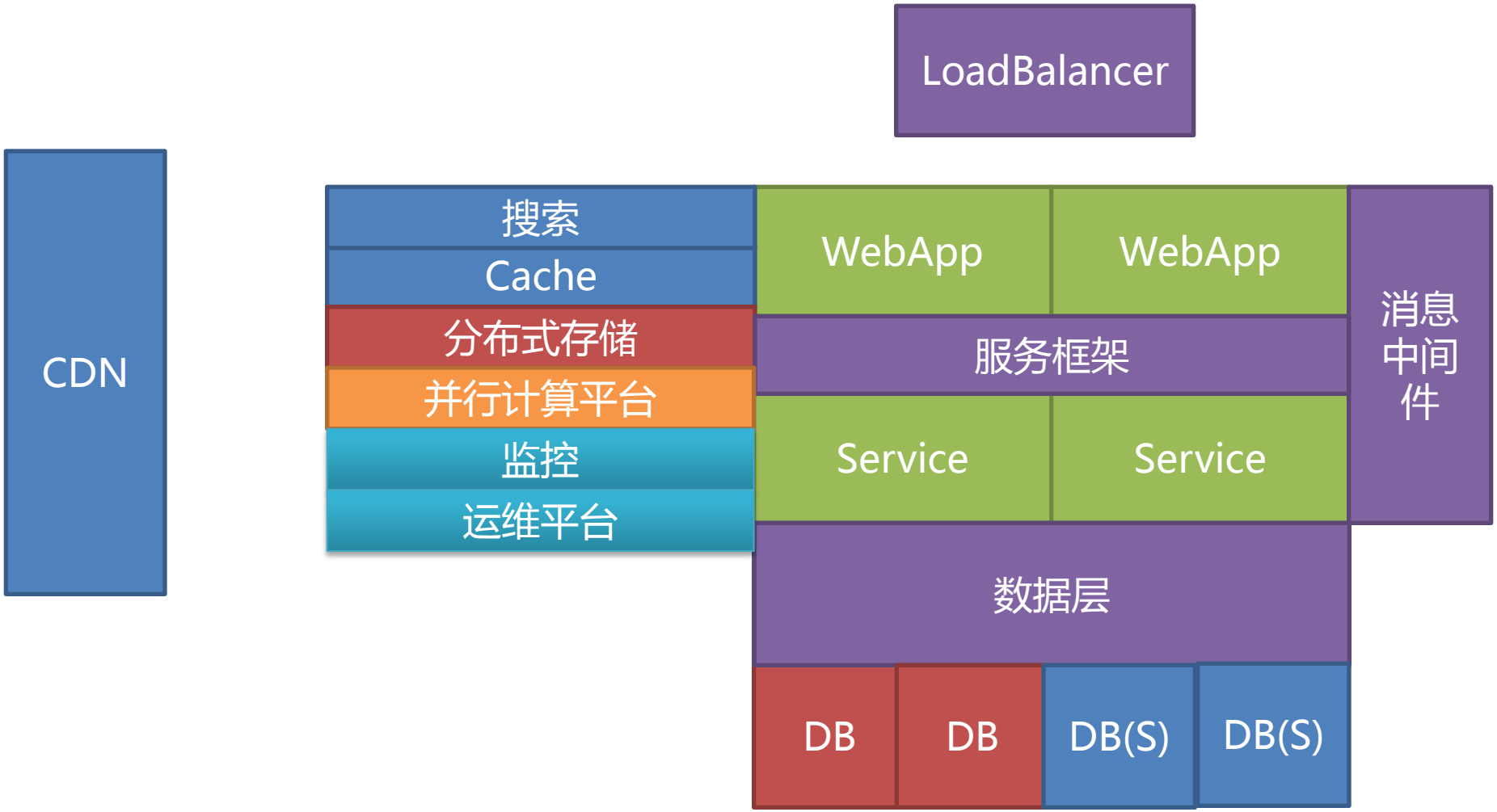

LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

分布式缓存

- Tair

- Redis

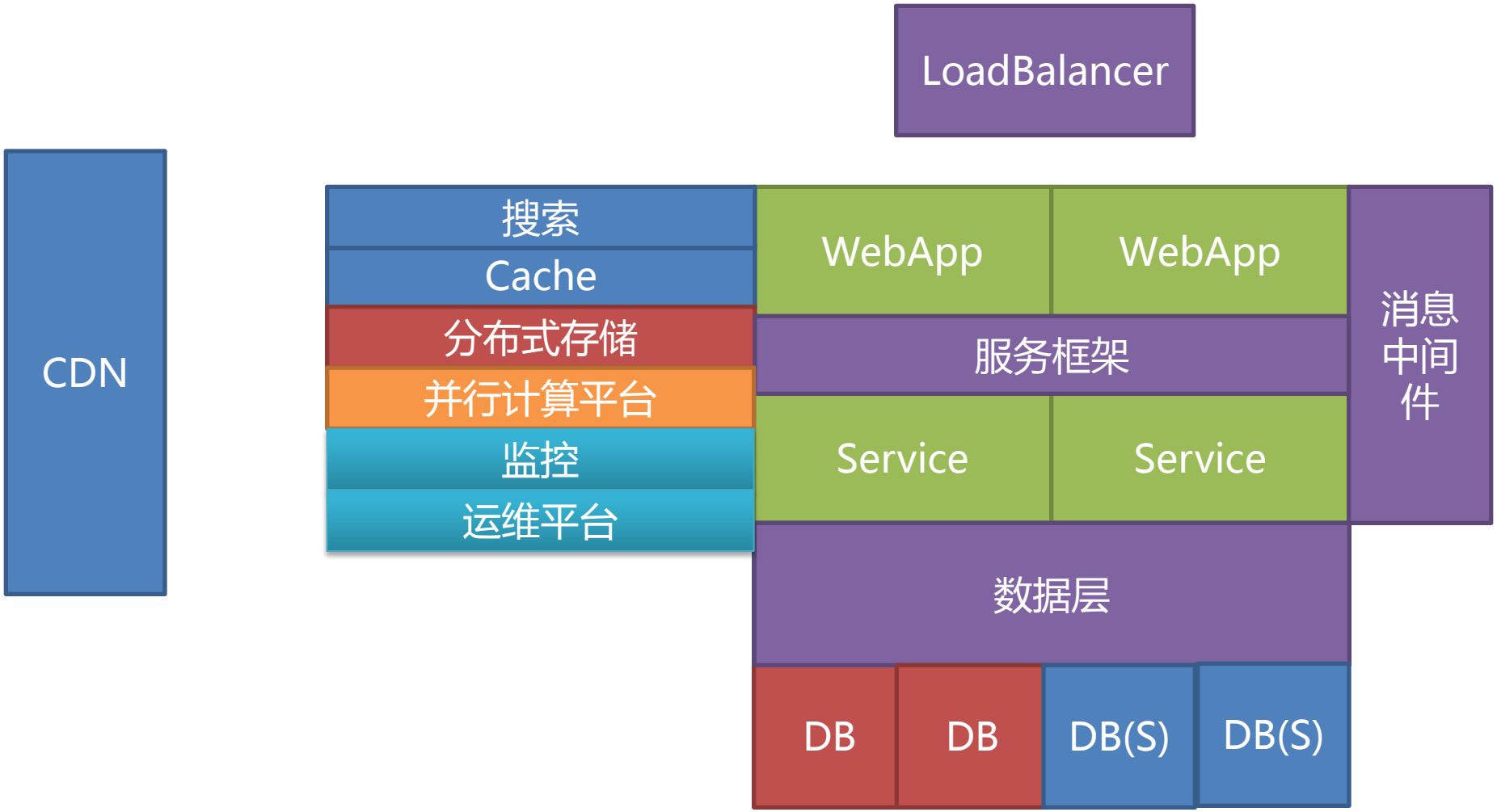

SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

分布式存储

- ✓TFS
- ✓OceanBase
- ✓Hbase
- ✓基于BDB的分布式存储


SALE

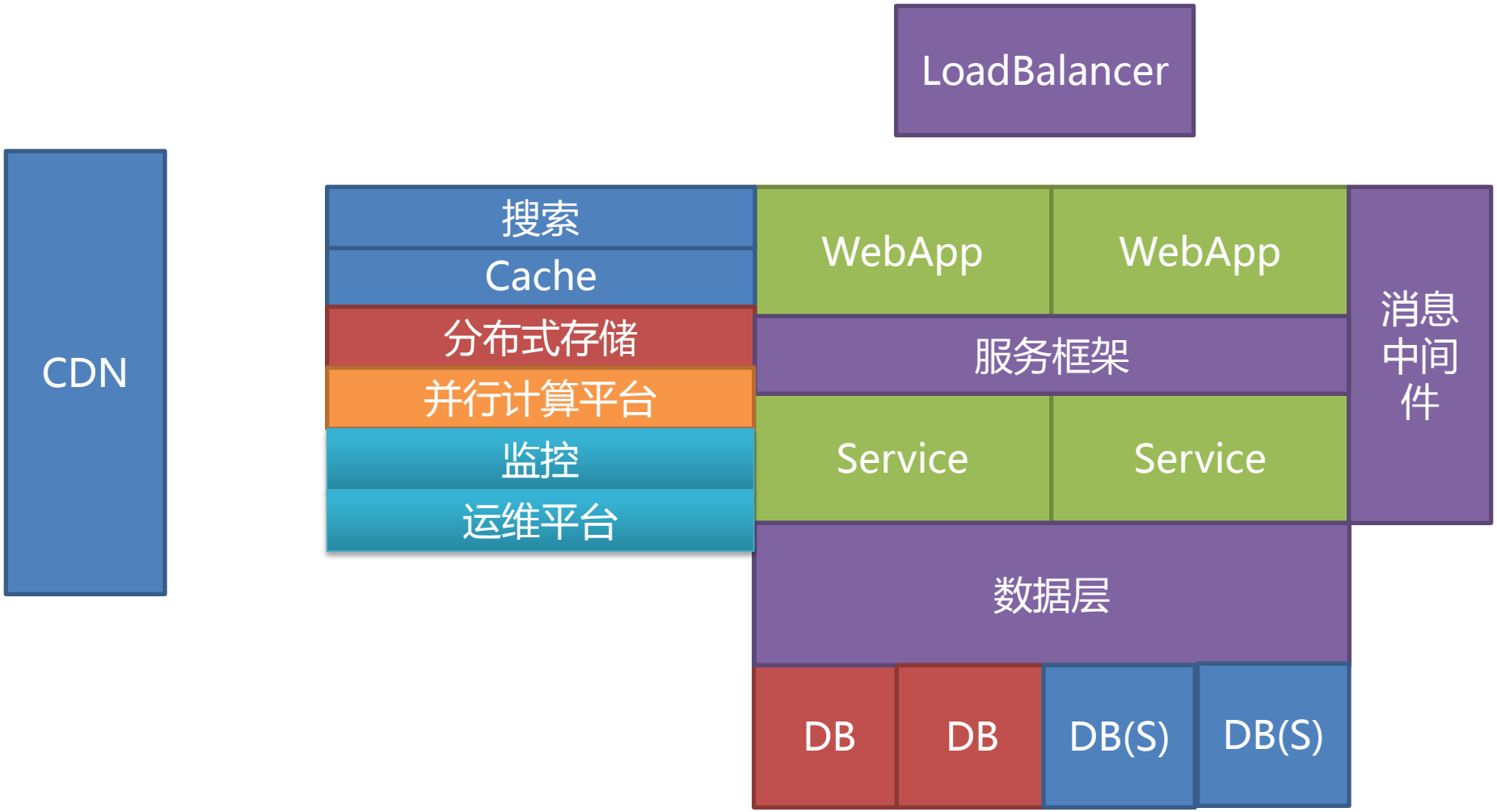

LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

搜索系统

- iSearch

- 基于Solr、Lucene的搜索系统


SALE

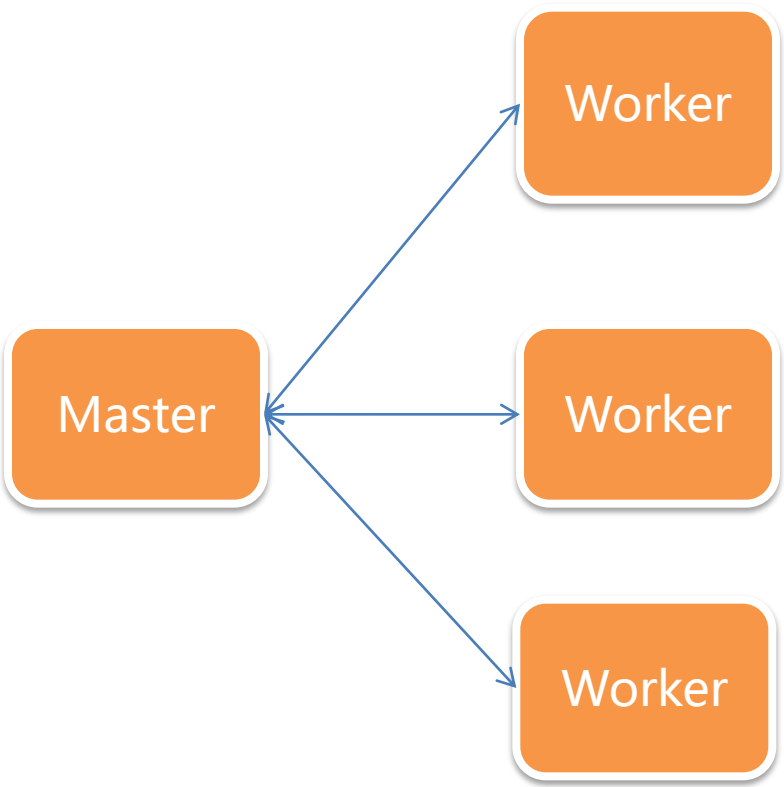

LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

Hadoop

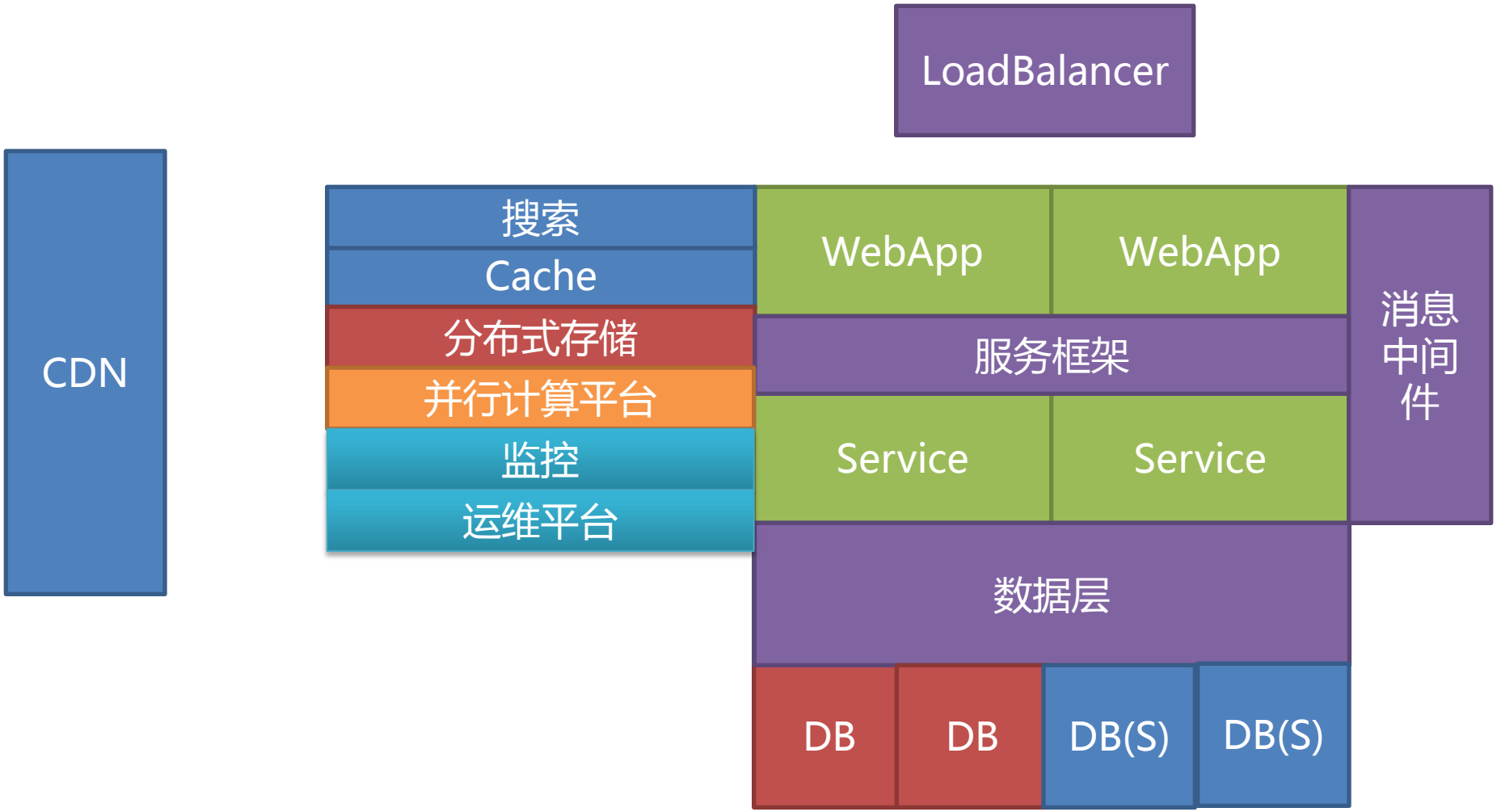
- 总容量25.35PB, 利用率55.2%
- 总共1400+台机器
- Master：8CPU(HT)，96GB内存，SAS Raid
- Slave节点异构
  - ✓ 8CPU/8CPU(HT)
  - ✓ 16G/24G内存
  - ✓ 1T x 12 / 2T x 6 / 1T x 6 SATA JBOD
  - ✓ 12/20 slots
- 约40000道作业/天, 扫描数据：约1.7PB/天
- 用户数474人, 用户组38个

FourInOne

- 简化版的实现
- 方便部署、维护和管理




SALE

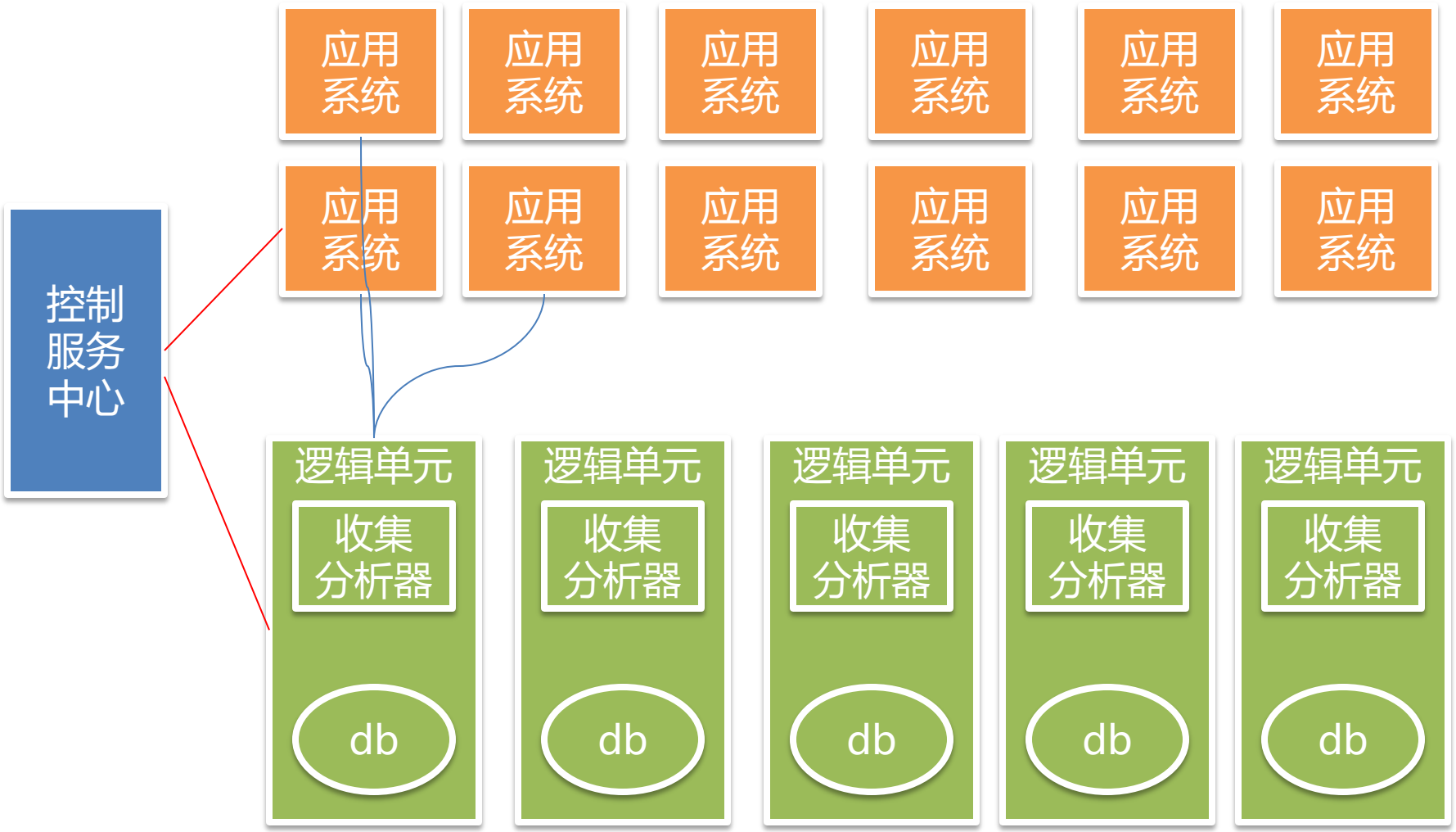

LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

CSP业务架构

依赖与降级	容量规划	运行状态(监控)
依赖关系系统	前台系统容量规划	中间件埋点监控模块(HSF、Tair、Search等)
稳定性开关模块(stable switch)	Center系统容量规划	系统信息监控模块(CPU,load,Memory)
强弱依赖验证系统	代码热点评估系统	应用Server监控模块(连接池，线程，内存)
自动/手动依赖降级	前端性能评估系统	Jvm、GC监控模块
	网络流量评估系统	系统日志监控模块
	容量预测/服务器预测	系统自定义埋点监控模块
		交易流程巡检(交易可用性)
		全量机器监控
		系统问题自动排查模块
稳定平台-各个业务线	定性平台-PE-性能测试-各个业务线	稳定平台-各个业务线
平台基础模块		
数据收集	报警	报表	备份	事件库	埋点/日志规范
控制中心		

CSP

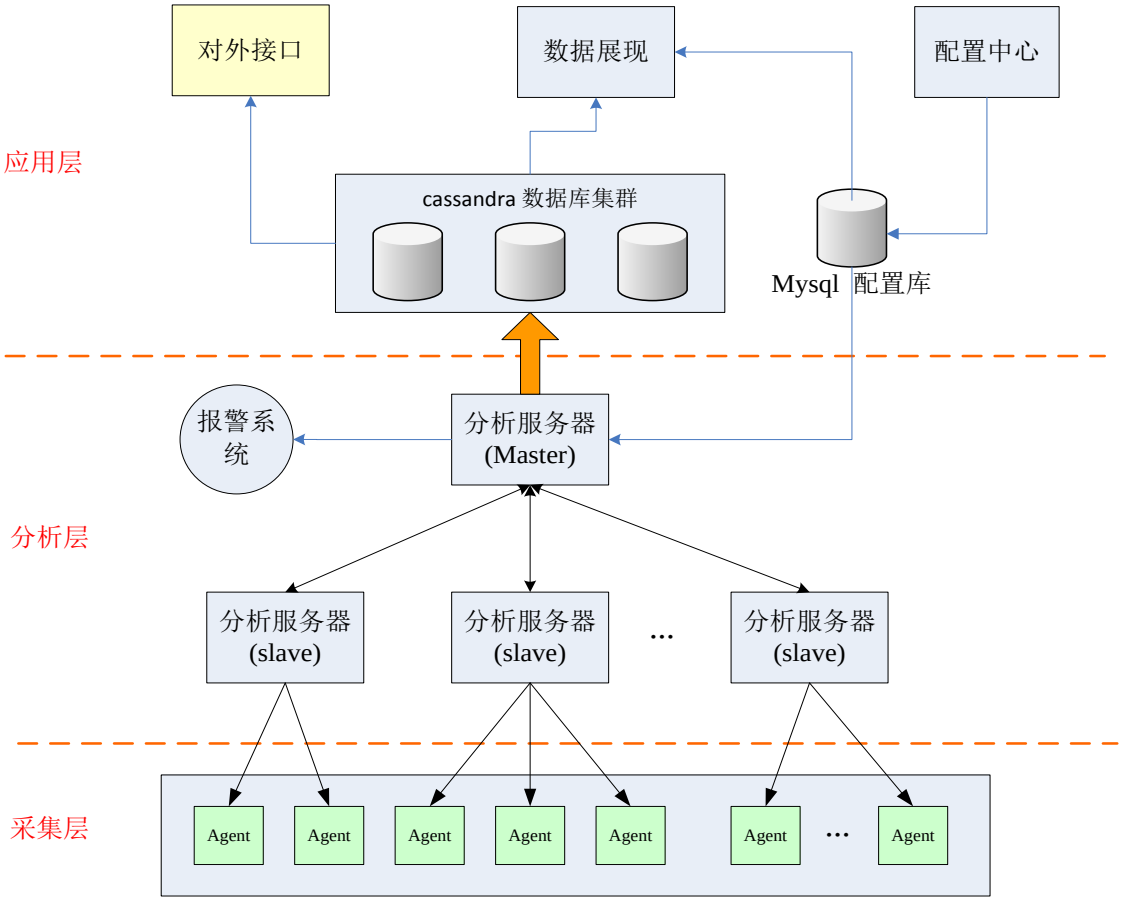

SALE


应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
应用系统
控制服务中心
逻辑单元
收集分析器
db
逻辑单元
收集分析器
db
逻辑单元
收集分析器
db
逻辑单元
收集分析器
db
逻辑单元
收集分析器
db

CSP的一些数据

- 平台化
  - ✓ 15台机器，15个库
  - ✓ APP采用了逻辑单元的概念(隔离)
  - ✓ 采用15个逻辑单元对超过200多个系统进行监控
- 每日数据超过5000万
- 24个系统进行容量规划
- 10个系统进行依赖关系
- 6个系统完成统一的保护开关部署

哈勃系统架构


对外接口
数据展现
配置中心
应用层
cassandra 数据库集群
Mysql 配置库
报警系统
分析服务器 (Master)
分析层
分析服务器 (slave)
分析服务器 (slave)
...
分析服务器 (slave)
采集层
Agent
Agent
Agent
Agent
Agent
Agent
...
Agent

系统特点

- 超轻量级HubAgent：
  - ✓ HubAgent是一个基于插件模式轻量级HTTP服务，可以通过增加插件的方式进行扩展，目前已提供“增量日志获取”，“外部程序（脚本）调用”，“进程&线程探测”，“端口探测”等多种服务。系统使用Python开发，不依赖第三方系统，可以安装在任何linux服务器上。
- 分布式的分析系统：.
  - ✓ 分析系统使用MapReduce的编程模型，对大量的监控数据进行并行处理。系统将采用松耦合的Master-Slave模式，Master处于被动状态，只负责任务的生成和合并，不需要关心任务分派。系统设计更加简单、灵活，通过增加Slave即可得到水平扩展能力。
- Cassandra数据存储：
  - ✓ 系统使用Cassandra来实现大量监控数据的分布式存储，Cassandra具有模式灵活，扩展性强，具有多维数据结构，支持范围查询等特点。而它“写入快，读取慢”的特性也恰好符合监控系统“写多读少”的特点，非常适合做监控数据的存储系统。

哈勃运行状况

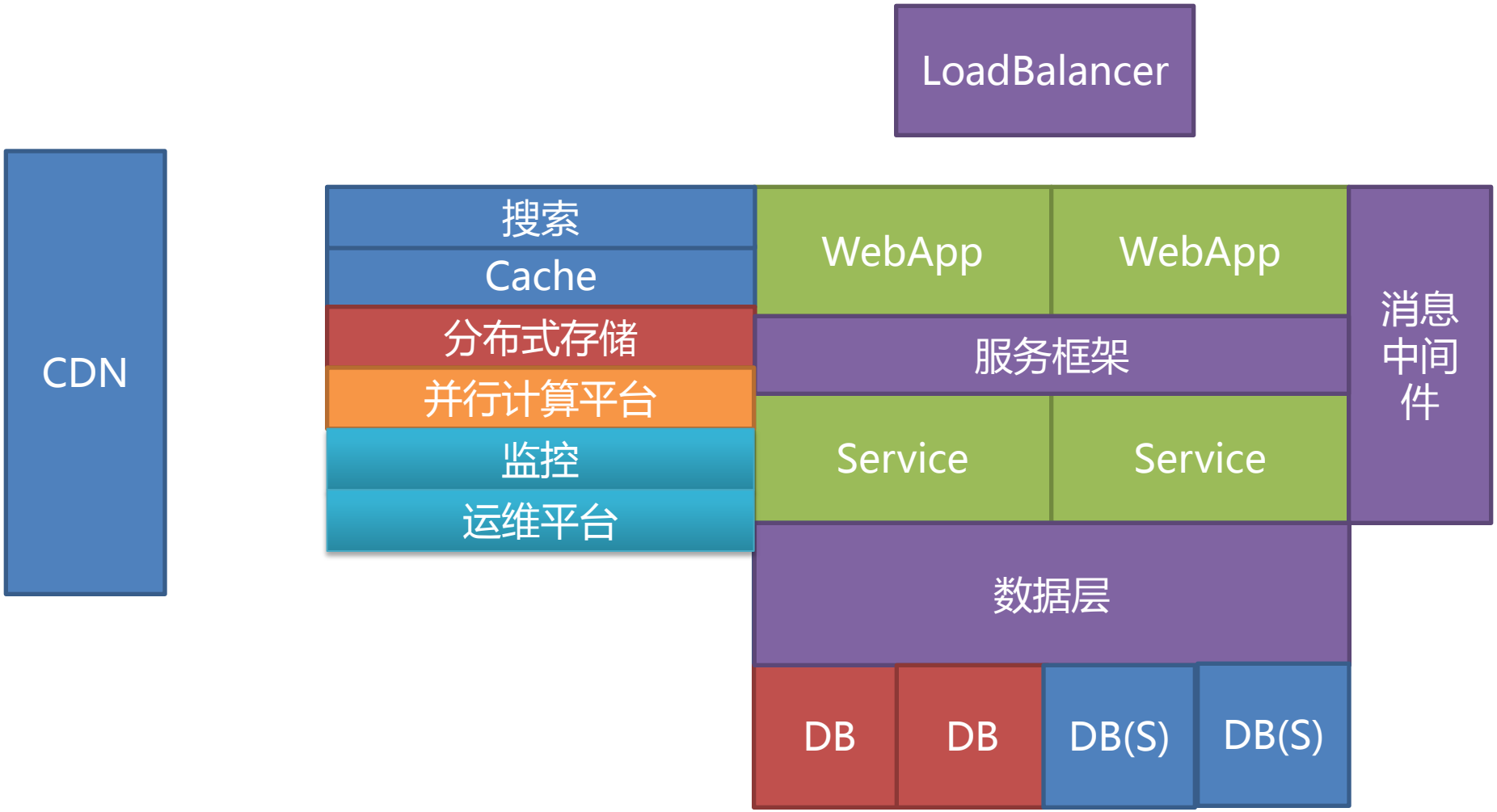
➢ 应用情况

- ✓ 目前以接入包括主站和广告在内的339 应用，配置了超过1500个任务和3000多张监控报表。能为应用的集群指标提供30天的明细数据，单机提供7天的明细数据；每小时的统计数据保存1年和每天的统计数据永久保存。

➢ 系统情况：

- ✓ HubAgent已经安装超过3500台，基本覆盖了主站的应用服务器。
- ✓ 目前有1台Master服务器和7台Slave服务器用于数据分析。 master的流量在5Mbps，Slave合计的平均网络流量60Mbps（峰值120Mbps），平均每天处理约4T的原始数据。
- ✓ 目前有5台Cassandra服务器，350G左右的数据存储量，每天的数据增量在25G左右；保存了超过18万个监控指标。

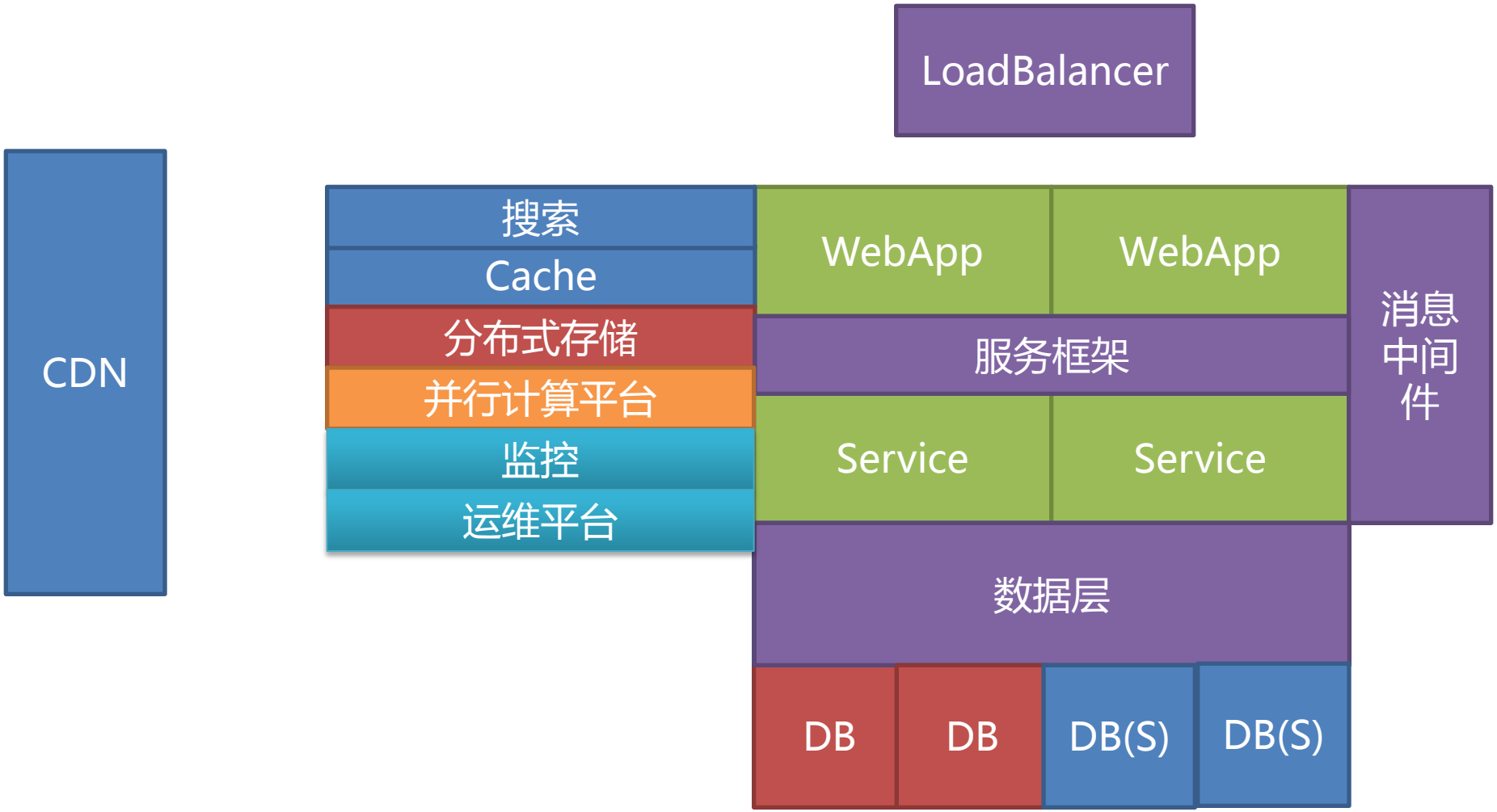

SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

运维平台

- 部署
- 发布
- 监控
- 管理


SALE


LoadBalancer
CDN
搜索
Cache
分布式存储
并行计算平台
监控
运维平台
WebApp
WebApp
服务框架
Service
Service
消息中间件
数据层
DB
DB
DB(S)
DB(S)

我们面临的挑战

TAOBAOJM

面临的挑战

稳定
99.99%?

运维
10w台机器/10个机房/300个应用？

安全
用户信息/钓鱼/攻击？

数据
PB数据的存储/检索/分析？

成本
机器/带宽？

性能
1秒？

Thanks !

TAOBAOJM