基于万节点Kubernetes支撑大规模云应用实践

刘超 网易蜂巢解决方案总架构师

关于我

我是谁

我是刘超，爱代码，爱开源
http://blog.csdn.net/popsuper1982

略懂Lucene, OpenStack, Docker, Mesos

Open DC/OS 社区贡献者

从哪里来

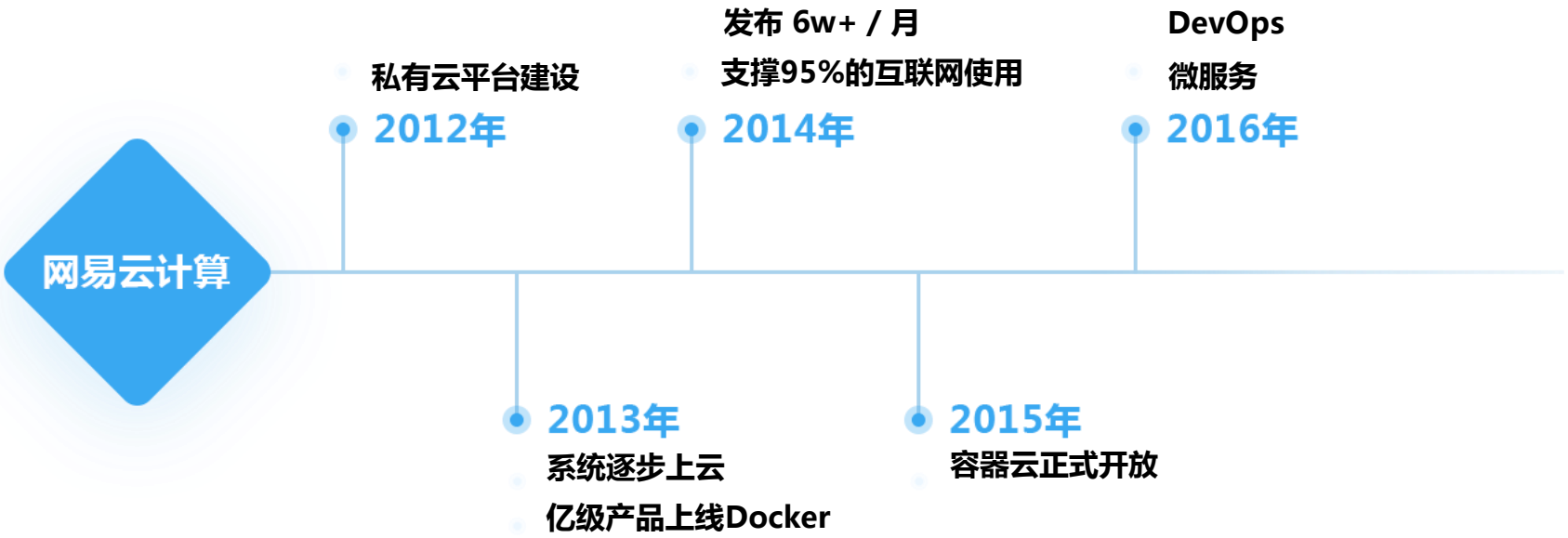
网易蜂巢解决方案总架构师

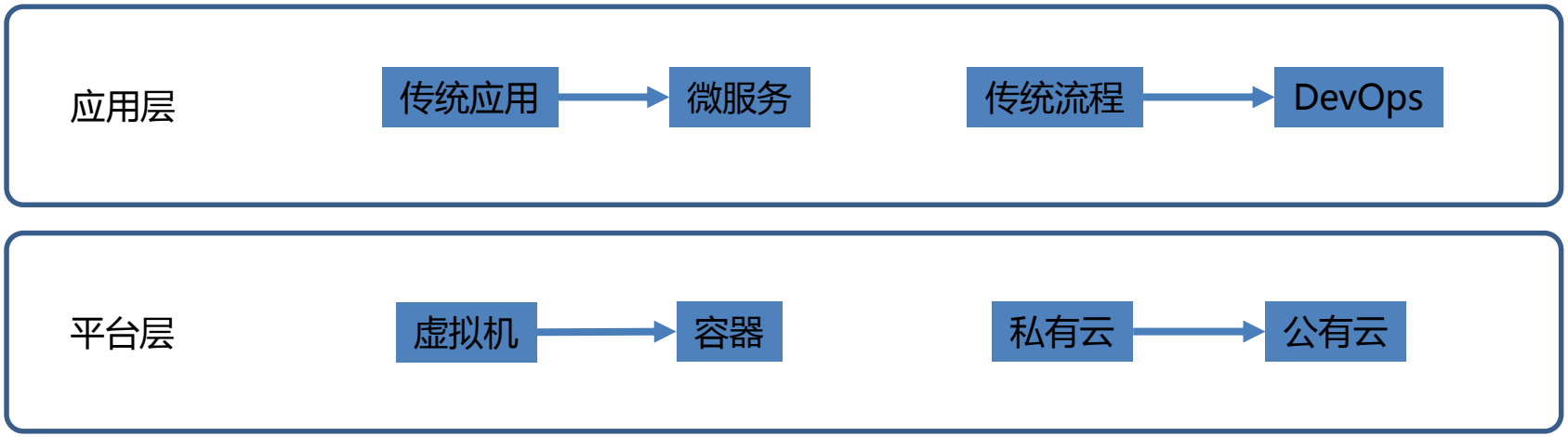
目标：写代码中最懂解决方案的，懂解决方案中最会写代码的

到哪里去

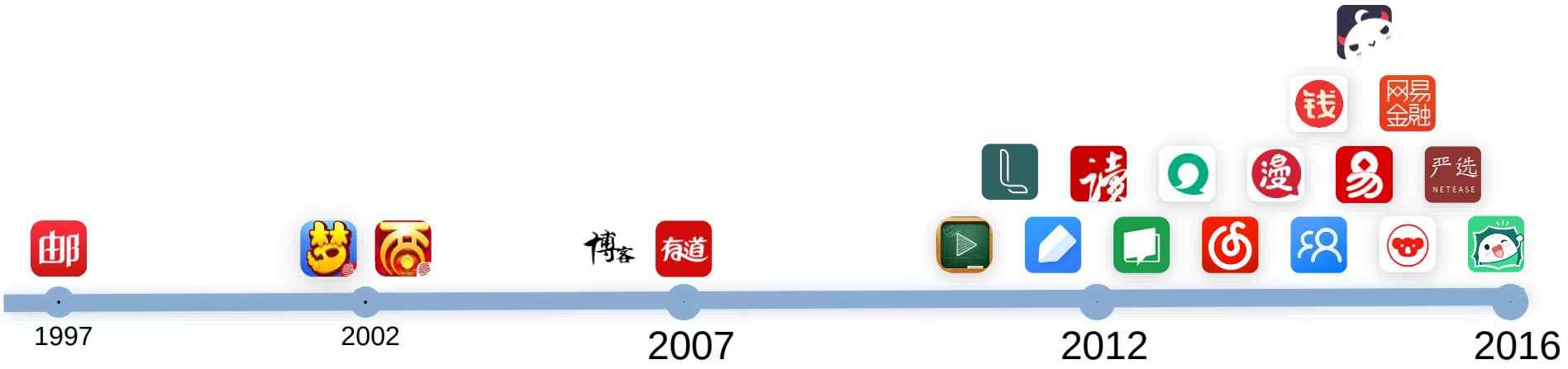
相信趋势：互联网+，容器，公有云，微服务，DevOps

网易蜂巢历程


网易云计算
私有云平台建设
2012年
2013年
系统逐步上云
亿级产品上线Docker
发布 6w+ / 月
支撑95%的互联网使用
2014年
2015年
容器云正式开放
DevOps
微服务
2016年

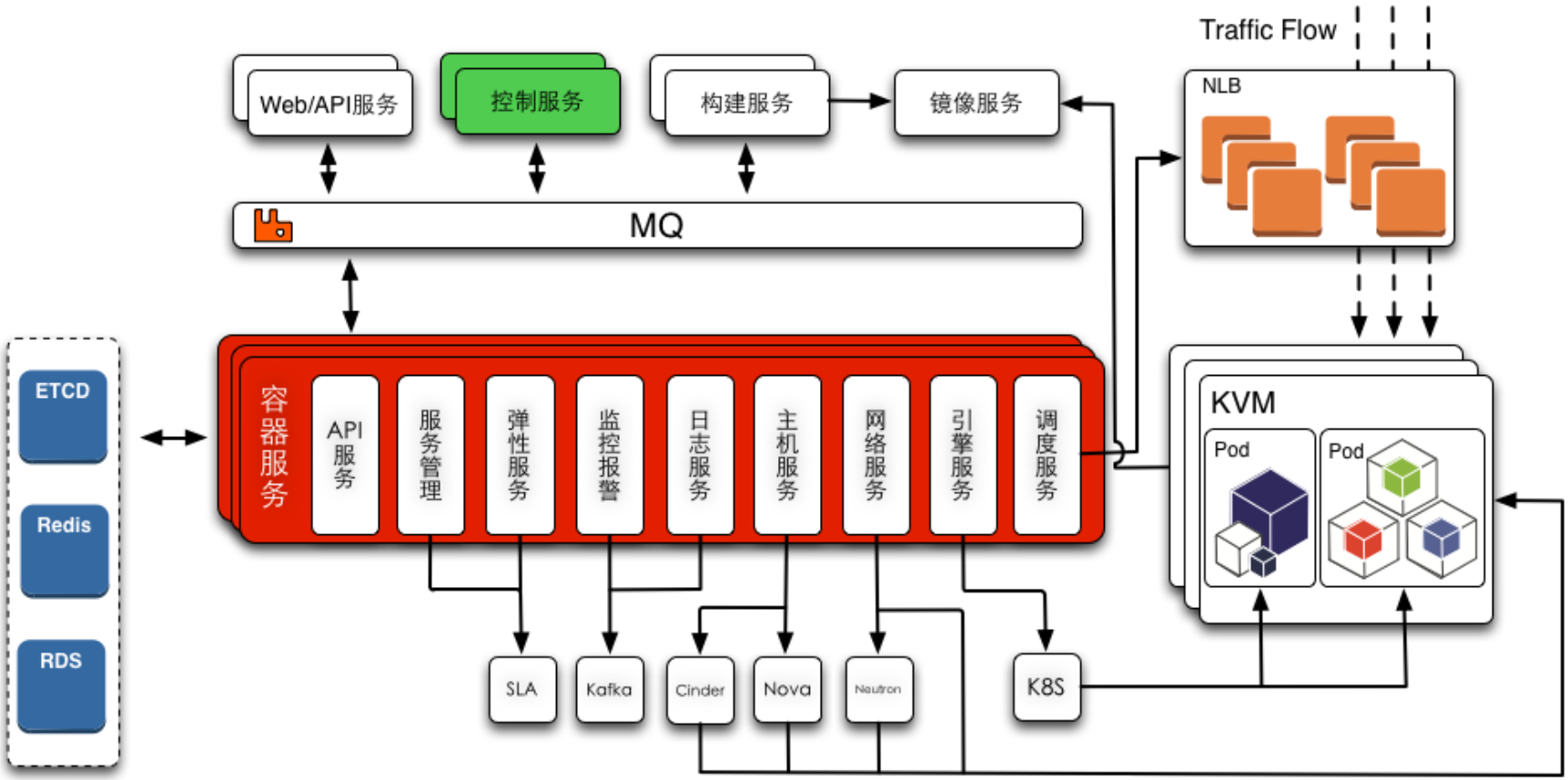

应用层
传统应用
微服务
传统流程
DevOps
平台层
虚拟机
容器
私有云
公有云

网易蜂巢上的大规模应用



6.18 **11.11** **200,000,000用户**

网易蜂巢的大规模容器平台



PART 02

私有云

私有云平台实现资源弹性

私有云平台优化

高可用，高性能PaaS

- 数据库：网易定制的MySQL内核分支，主从切换数据零丢失，提供健康检查和SQL优化工具
- 缓存服务：主从热备、跨可用域部署，自动容灾，高性能单笔延时毫秒级
- 对象存储：高可用性为99.99%，高可靠性三备份8个9，基于自研分布式非结构化存储系统

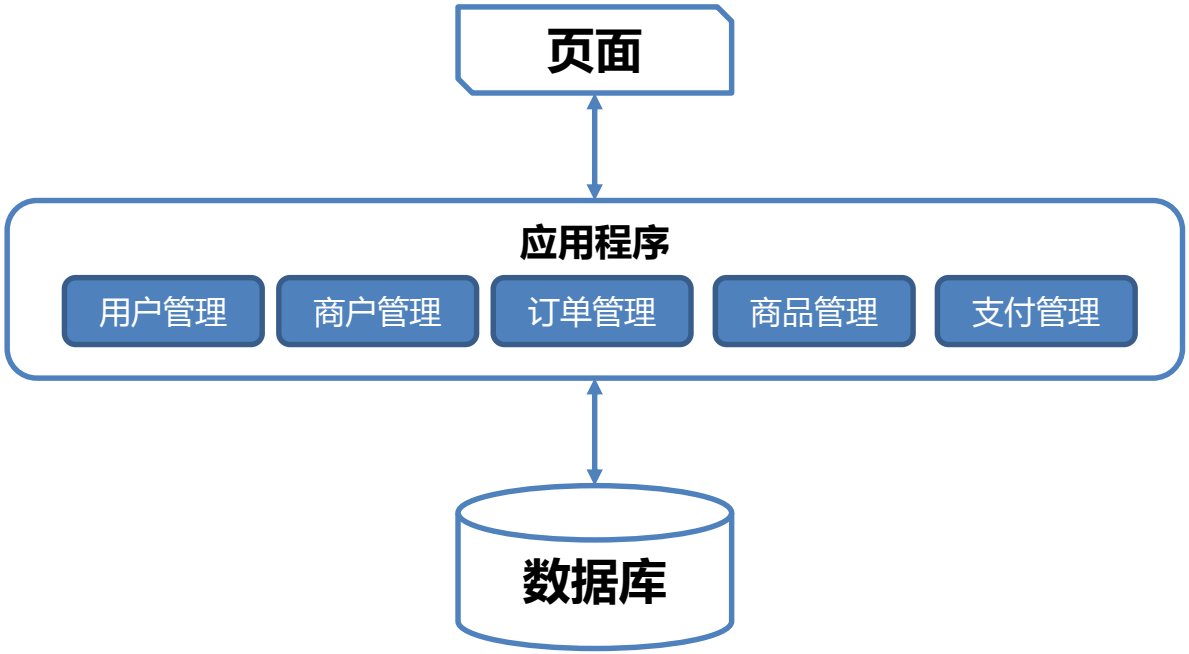
基于OpenStack自研IaaS

- 计算：定制KVM系统镜像，实现云主机IP静态化，优化OpenStack创建云主机流程
- 网络：二层至四层网络过滤防止MAC/IP欺骗，基于Linux TC修改OVS实现网络QoS
- 存储：云硬盘架构基于iscsi和Ceph实现，优化Ceph核心模块OSD

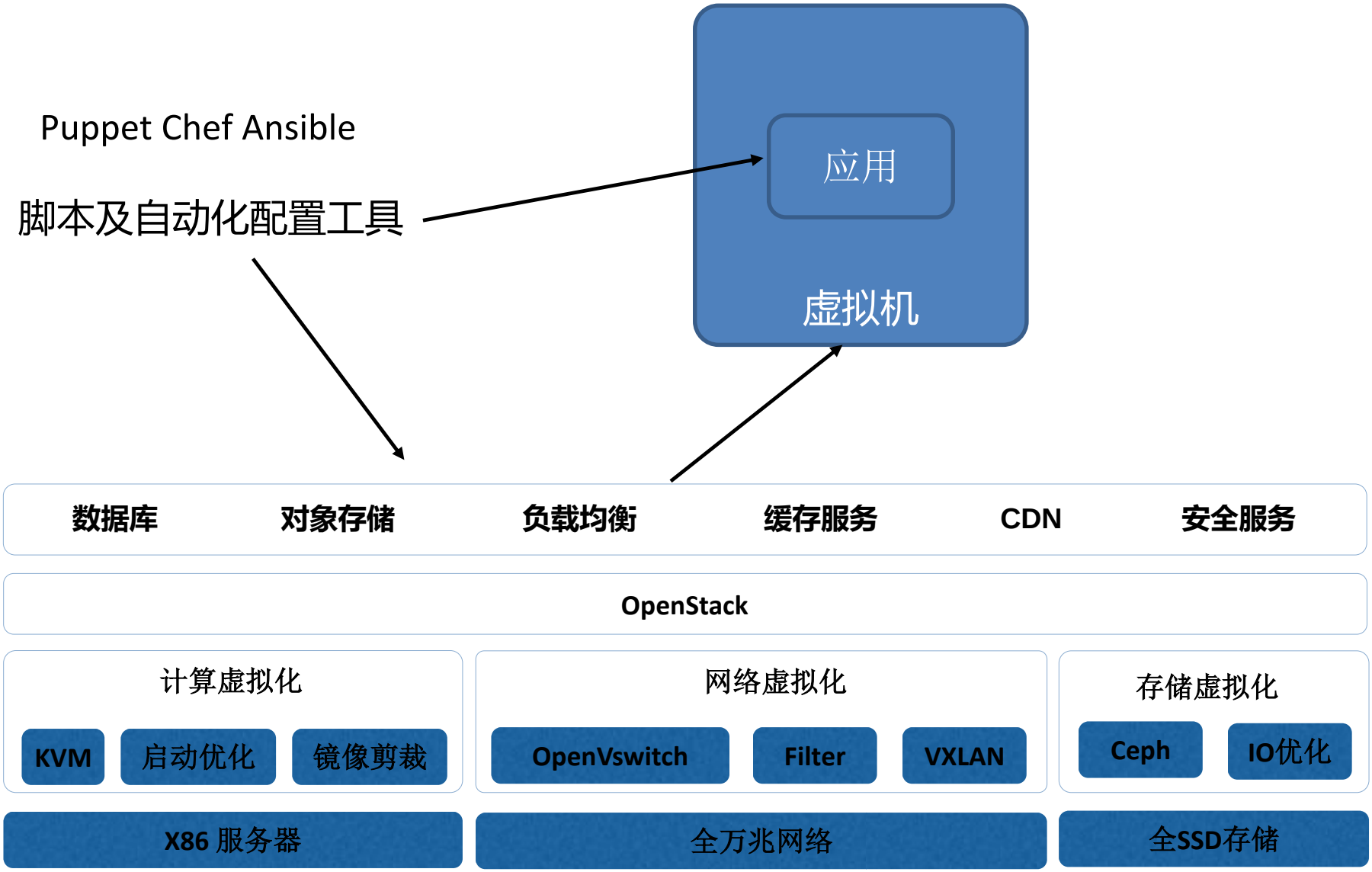
高规格硬件设备

- 多线BGP网络接入，万兆网络互联，全SSD存储

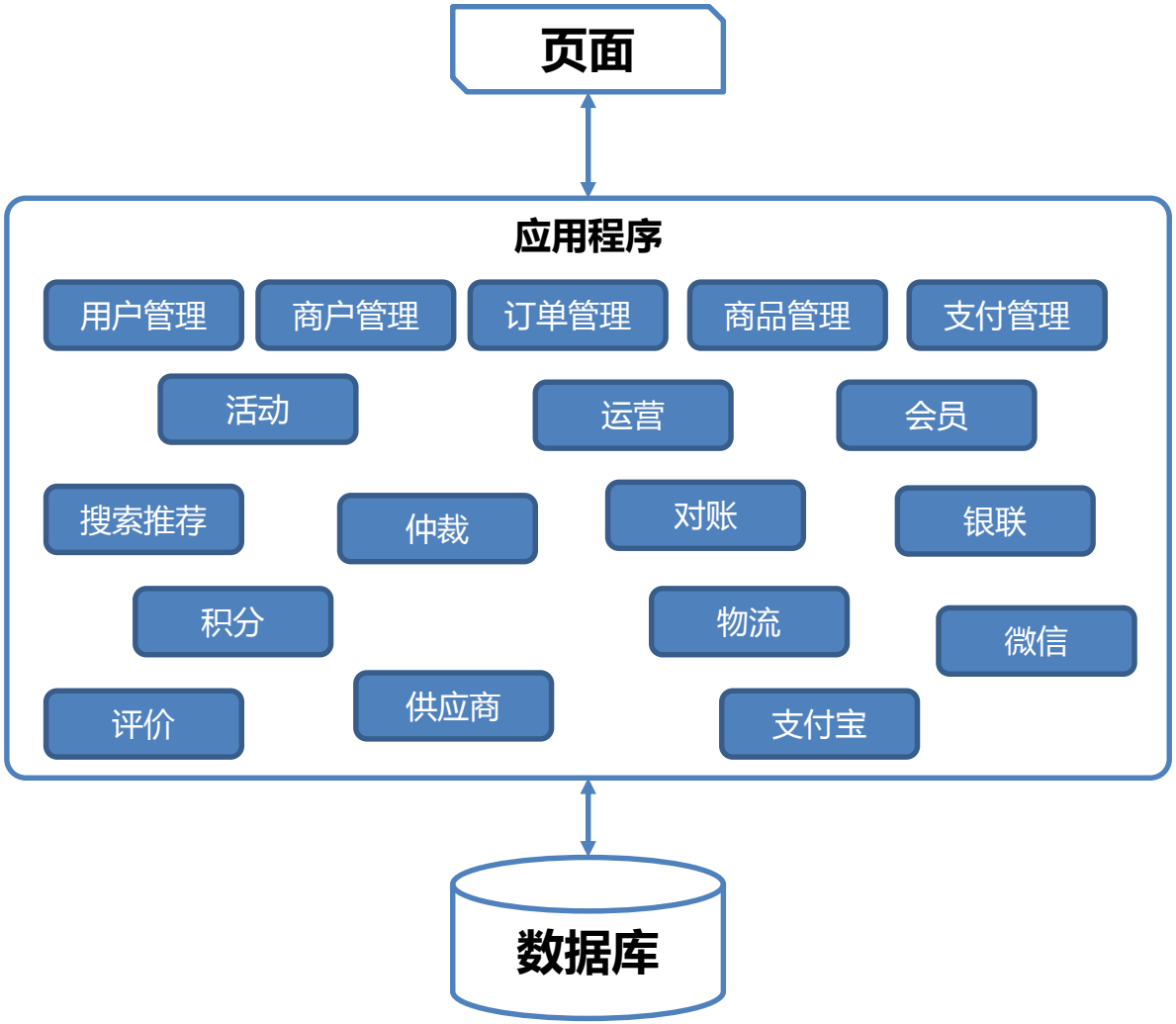
应用层架构雏形



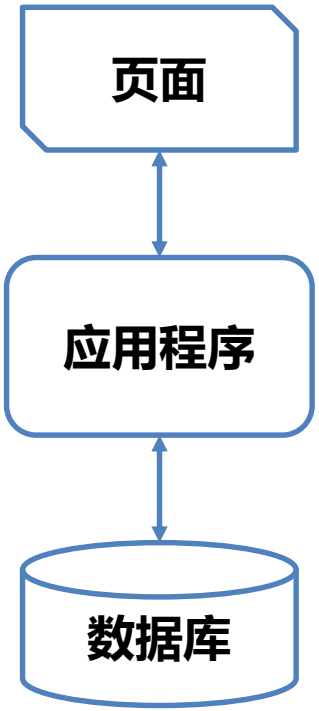
应用层虚机部署


Puppet Chef Ansible
脚本及自动化配置工具
应用
虚拟机
数据库
对象存储
负载均衡
缓存服务
CDN
安全服务
OpenStack
计算虚拟化
KVM
启动优化
镜像剪裁
网络虚拟化
OpenVswitch
Filter
VXLAN
存储虚拟化
Ceph
IO优化
X86 服务器
全万兆网络
全SSD存储

应用层架构复杂化


页面
应用程序
用户管理
商户管理
订单管理
商品管理
支付管理
活动
运营
会员
搜索推荐
仲裁
对账
银联
积分
物流
微信
评价
供应商
支付宝
数据库

架构之痛



时间灵活性：

应用快速迭代，缩短客户需求到产品上线的时间

空间灵活性：

应用弹性伸缩，应对业务量突然增长后较短时间恢复

管理灵活性：

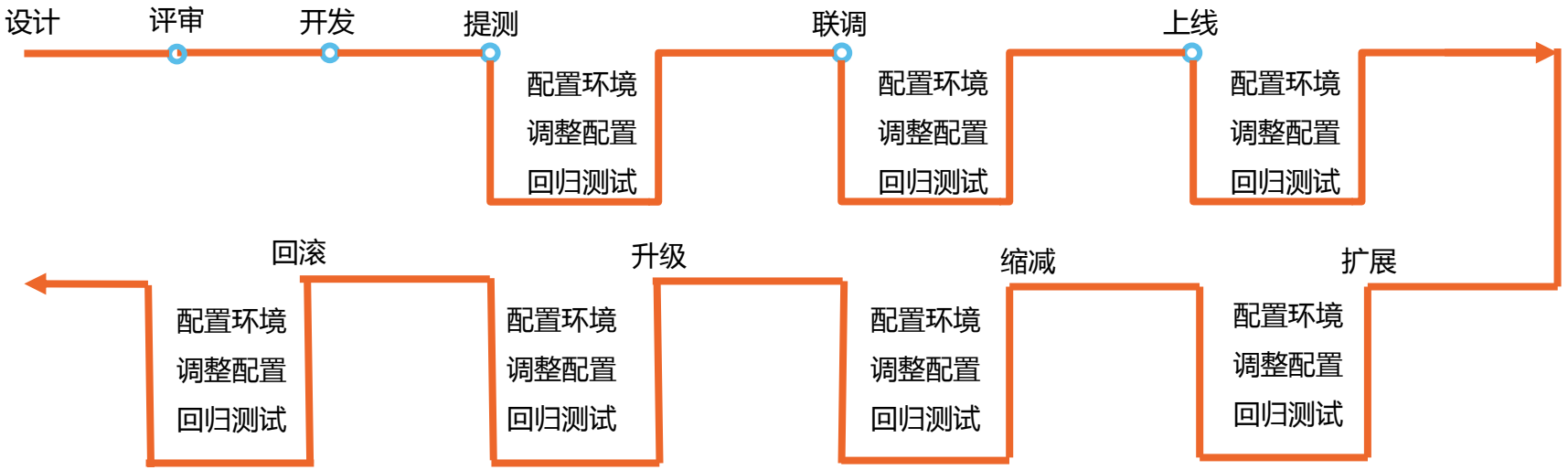
易部署，易迁移，服务发现，依赖管理，自动修复，负载均衡

时间灵活性

需求：策划一个营销活动，快速开发，快速部署，快速上线

现实：从开发Dev到运维Ops需要长长的流程

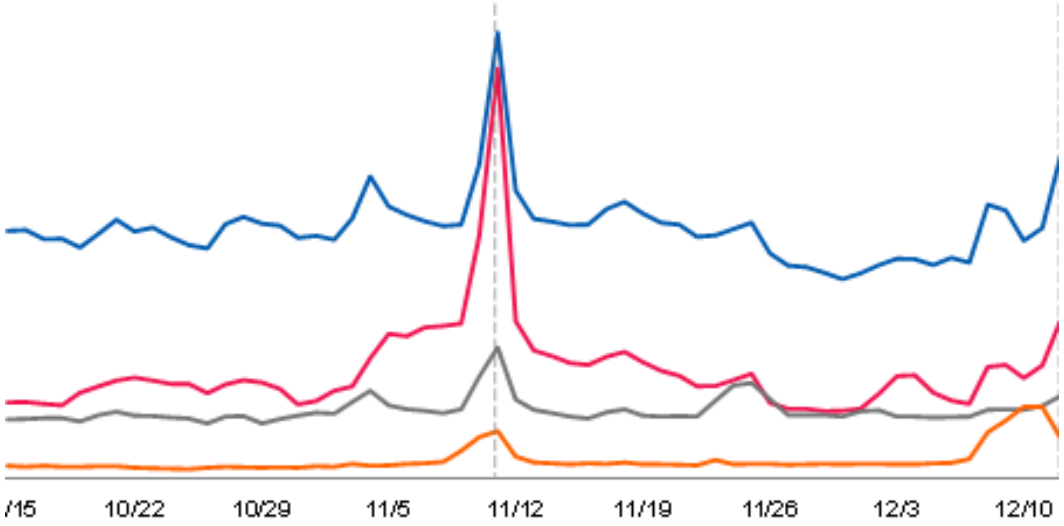
开发(Dev)：代码修改牵一发动全身



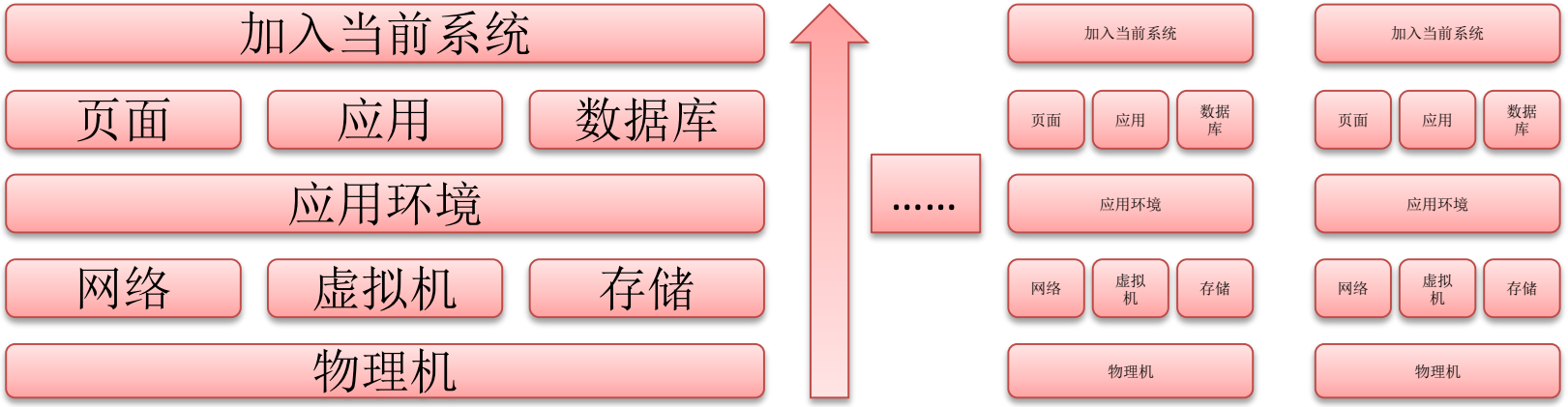
运维(Ops)：反复的部署，无法保证环境的一致

空间灵活性

需求：访问遭遇突发峰值，应用应该快速扩展提供支撑

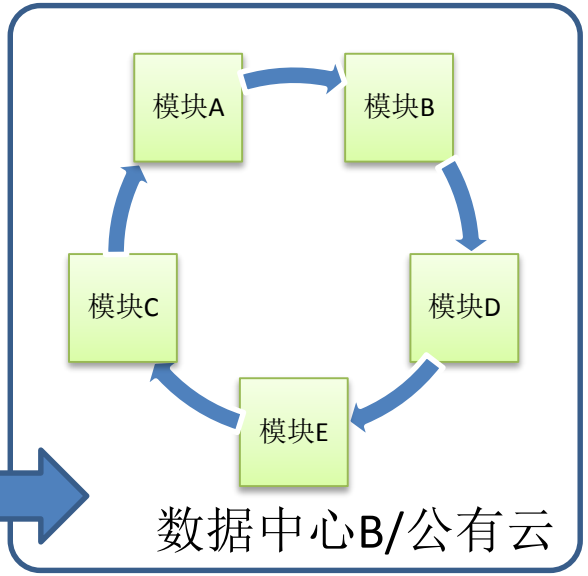


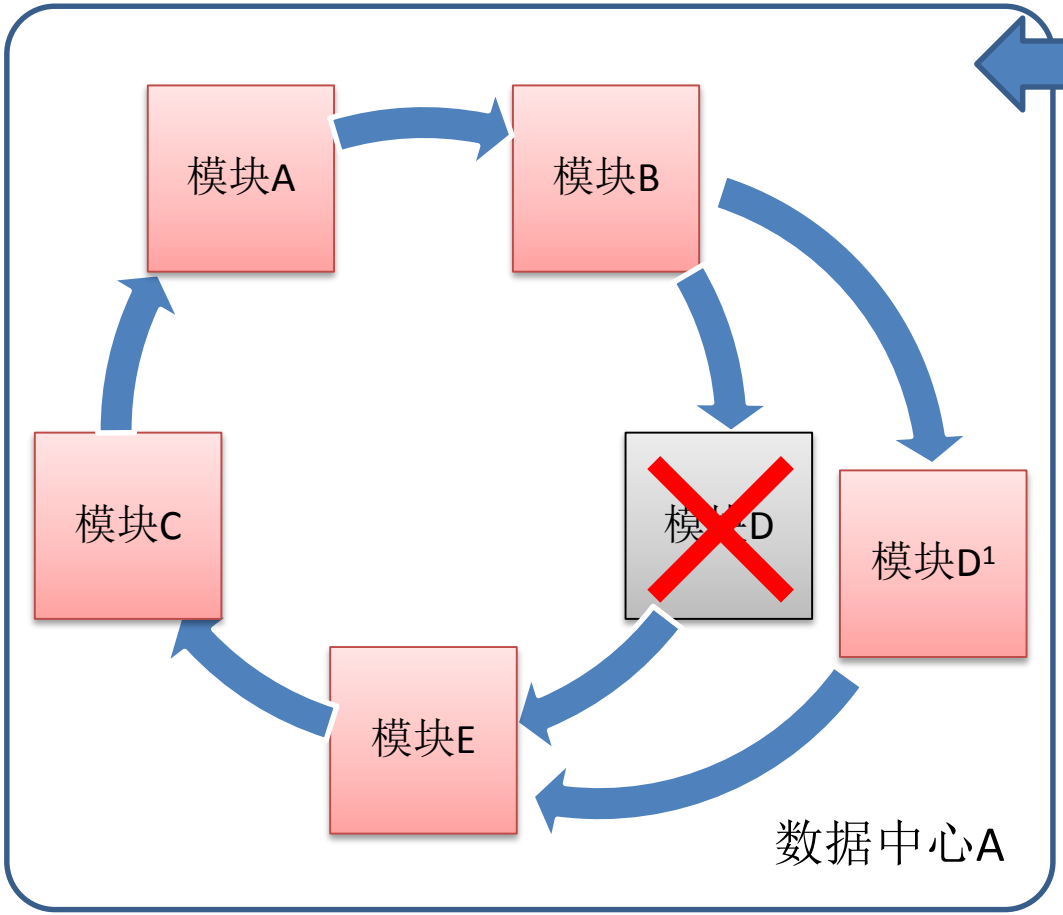
现实：万丈高楼平地起，一层一层慢慢盖



管理灵活性

需求：高可用，跨机房迁移，自动修复





现实：手动修复，手动迁移

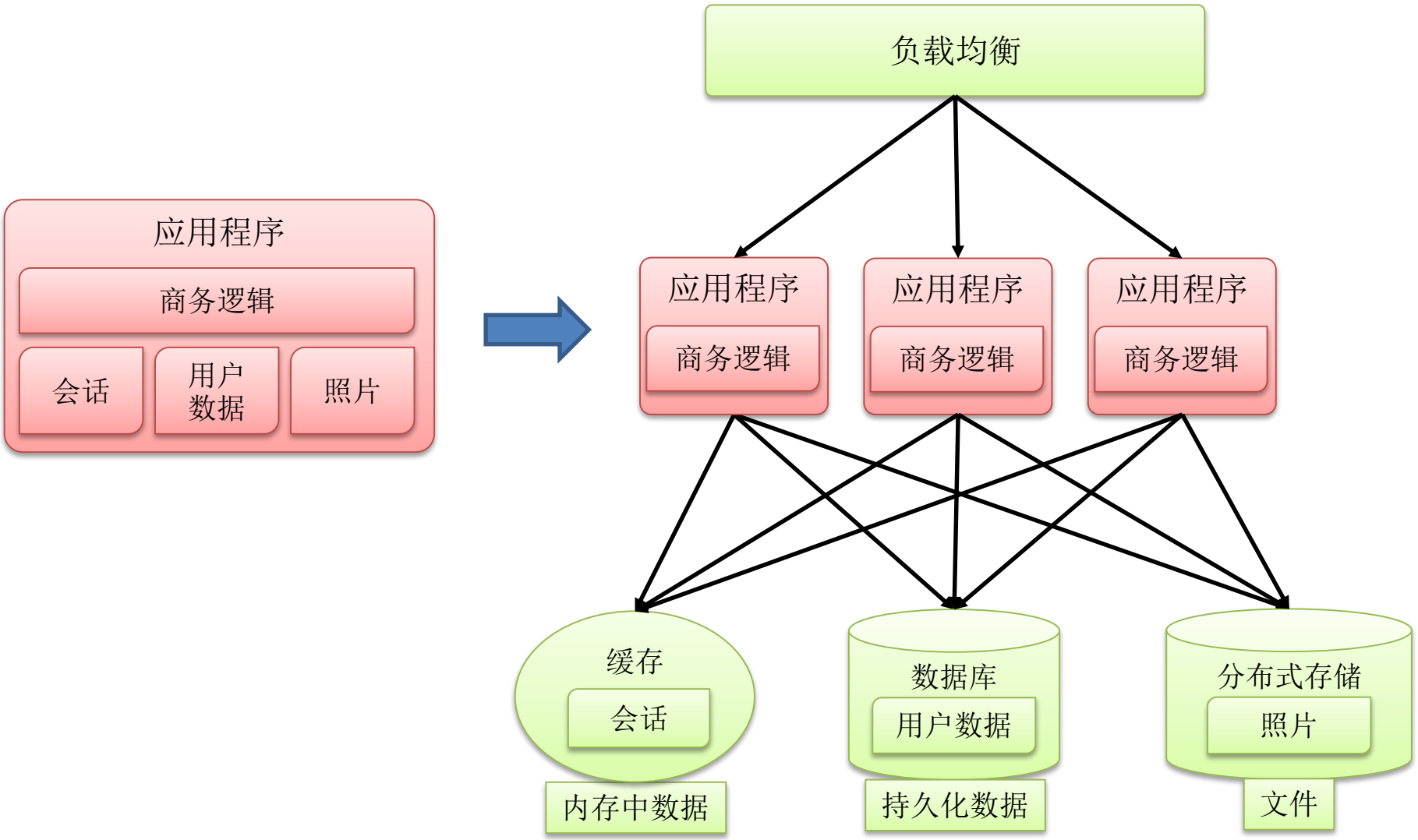
PART 03

容器云

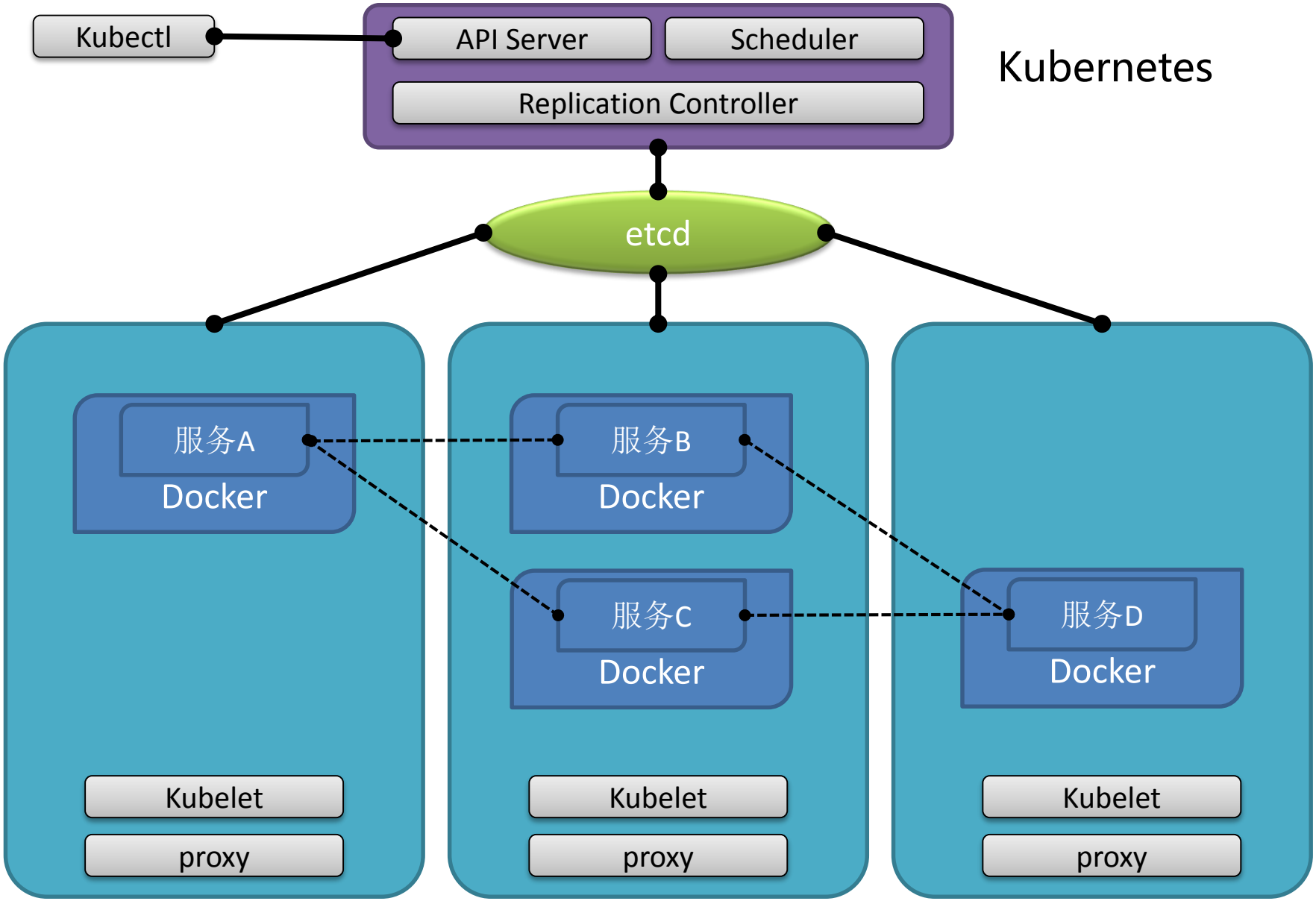
从虚拟机到容器

- 以资源为核心 -> 以应用为核心
- 有状态容器
- 容器跨主机互联
- 容器使用云盘存储

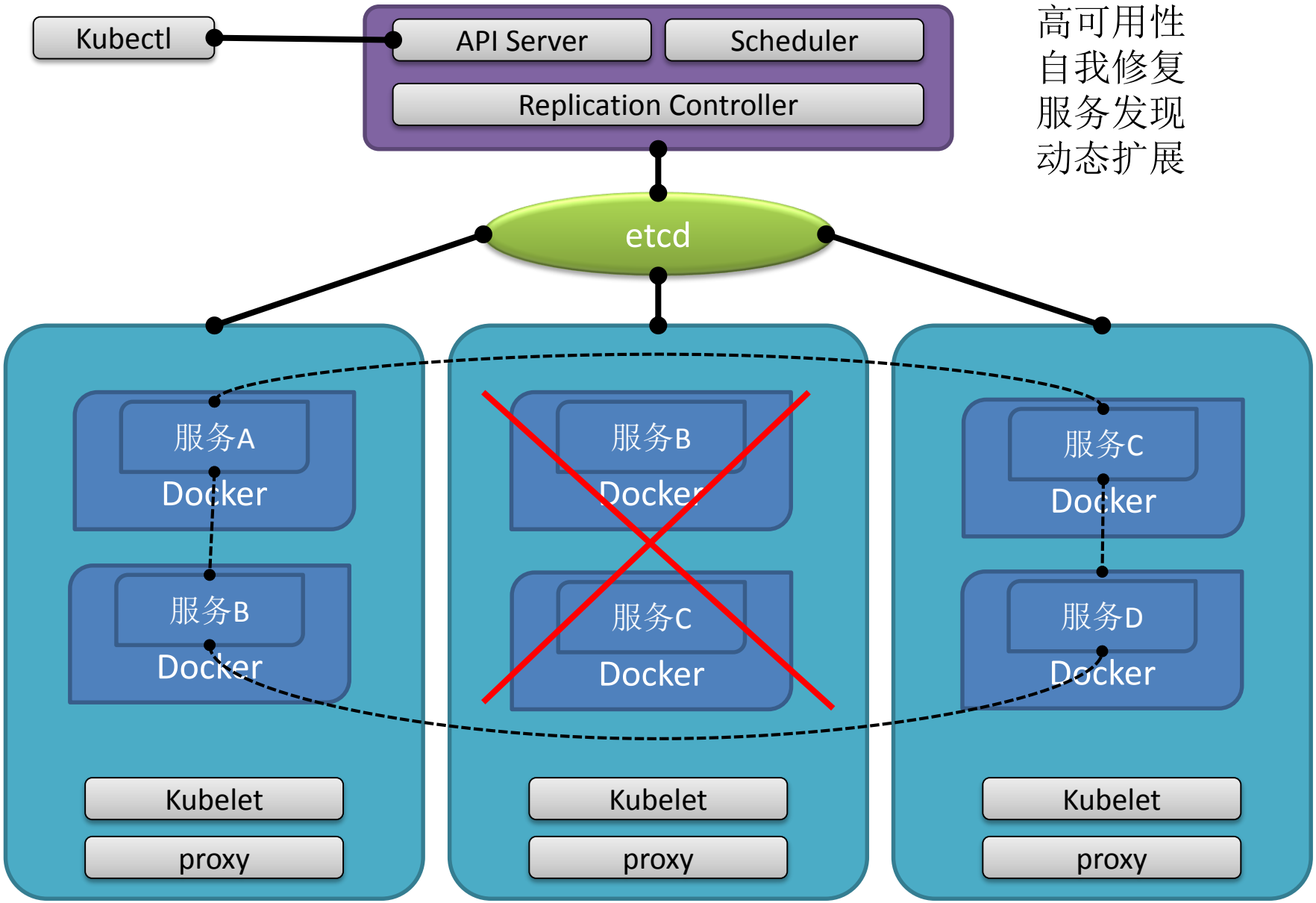
一板斧： 去状态、可扩展


应用程序
商务逻辑
会话
用户数据
照片
负载均衡
应用程序
商务逻辑
应用程序
商务逻辑
应用程序
商务逻辑
缓存
会话
内存中数据
数据库
用户数据
持久化数据
分布式存储
照片
文件

二板斧：容器化、可编排

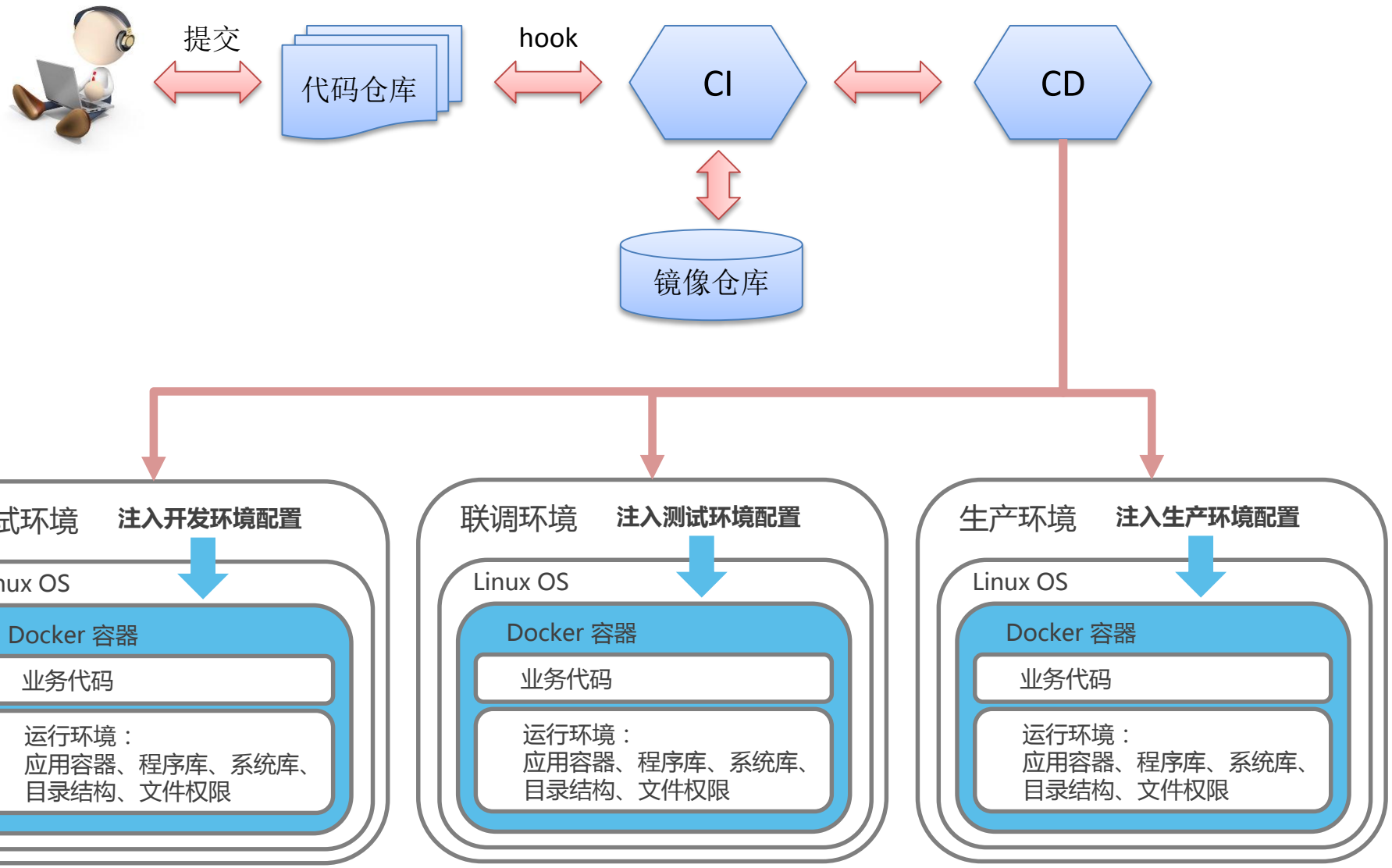

Kubectl
API Server
Scheduler
Replication Controller
Kubernetes
etcd
服务A
Docker
服务B
Docker
服务C
Docker
服务D
Docker
Kubelet
proxy
Kubelet
proxy
Kubelet
proxy

二板斧：容器化、可编排

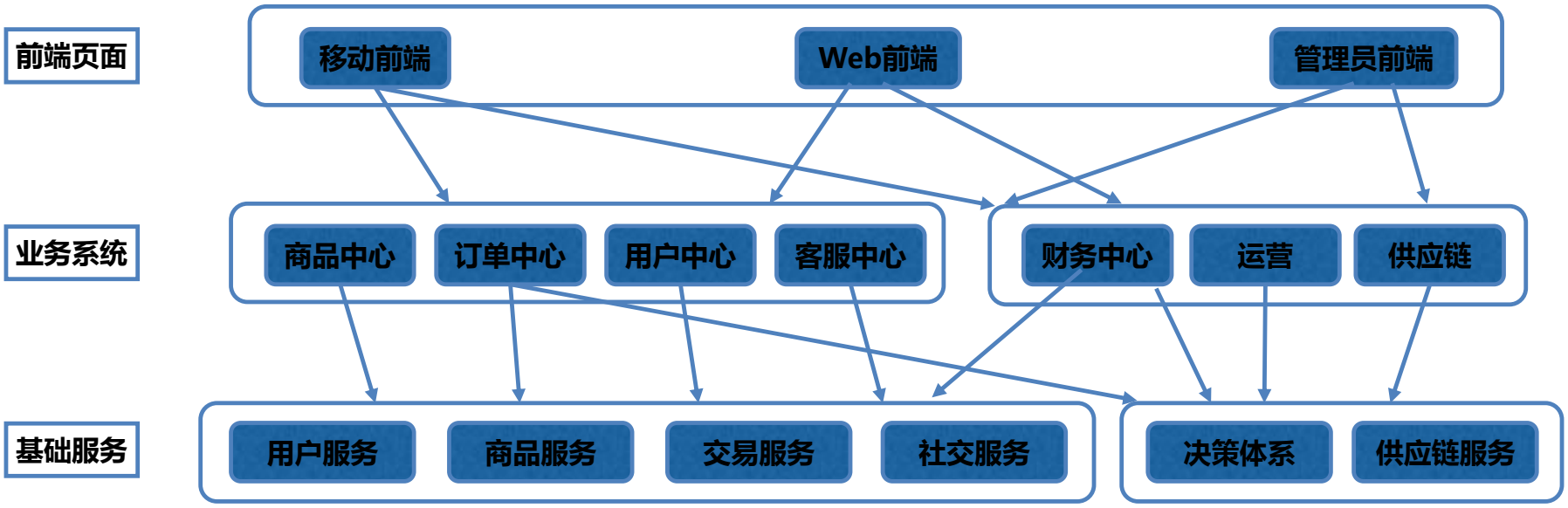


高可用性
自我修复
服务发现
动态扩展

三板斧：DevOps、可迭代



微服务架构


前端页面
移动前端
Web前端
管理员前端
业务系统
商品中心
订单中心
用户中心
客服中心
财务中心
运营
供应链
基础服务
用户服务
商品服务
交易服务
社交服务
决策体系
供应链服务

从私有云到公有云

- 容器的安全
- 容器的启动速度
- 容器的规模
- 容器的租户隔离

网易蜂巢平台


编排
有状态容器
调度优化
多租户
弹性伸缩
服务发现
容器
镜像仓库
镜像构建
日志服务
监控服务
持续集成
数据库
对象存储
负载均衡
缓存服务
CDN
安全服务
OpenStack
计算虚拟化
网络虚拟化
存储虚拟化

网易蜂巢平台

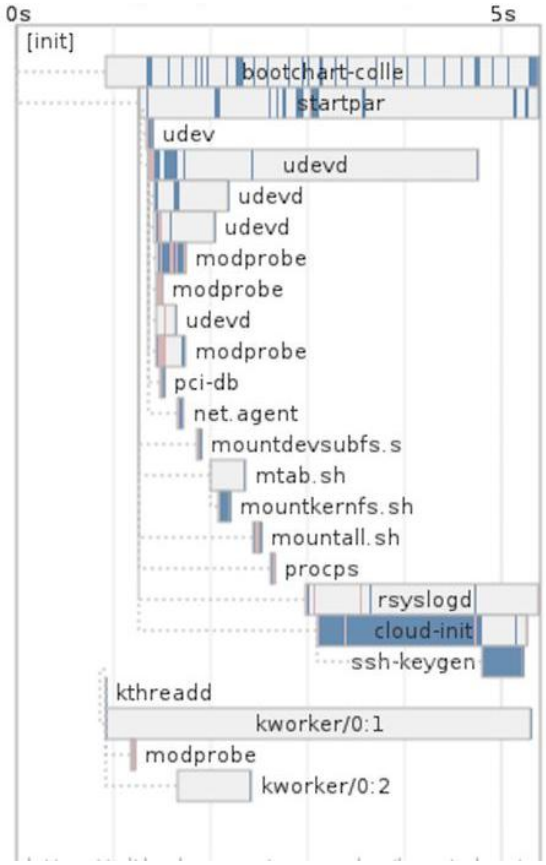
编排的优化

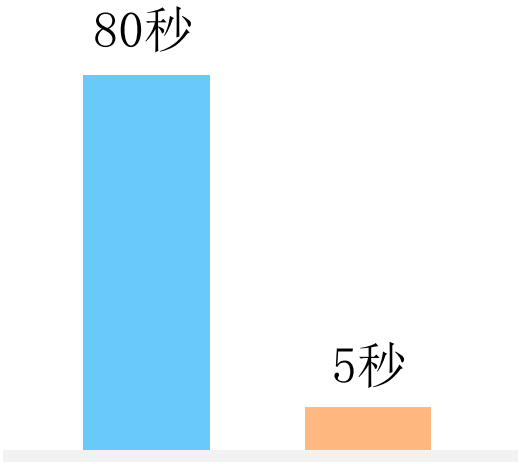
- 支持多租户: 默认kubernates的namespace只隔离replication controller，pod 等资源，网易实现节点，存储、网络的租户隔离
- 调度性能优化：kubernetes调度优化，任务串行队列改为多个优先级队列
- 集群扩展性：根据Pod/Node/Replication Controller等资源到拆分不同的etcd集群

容器的优化

- 虚拟化扁平二层网络，基于VXLAN实现租户隔离，外网网卡直接挂载到容器内部
- 有状态容器挂载云盘，可实现跨主机迁移
- 提供统一的日志收集，分析，搜索服务，利于分布式架构问题定位
- 引入服务端 APM 解决细粒度性能分析，迅速发掘性能瓶颈

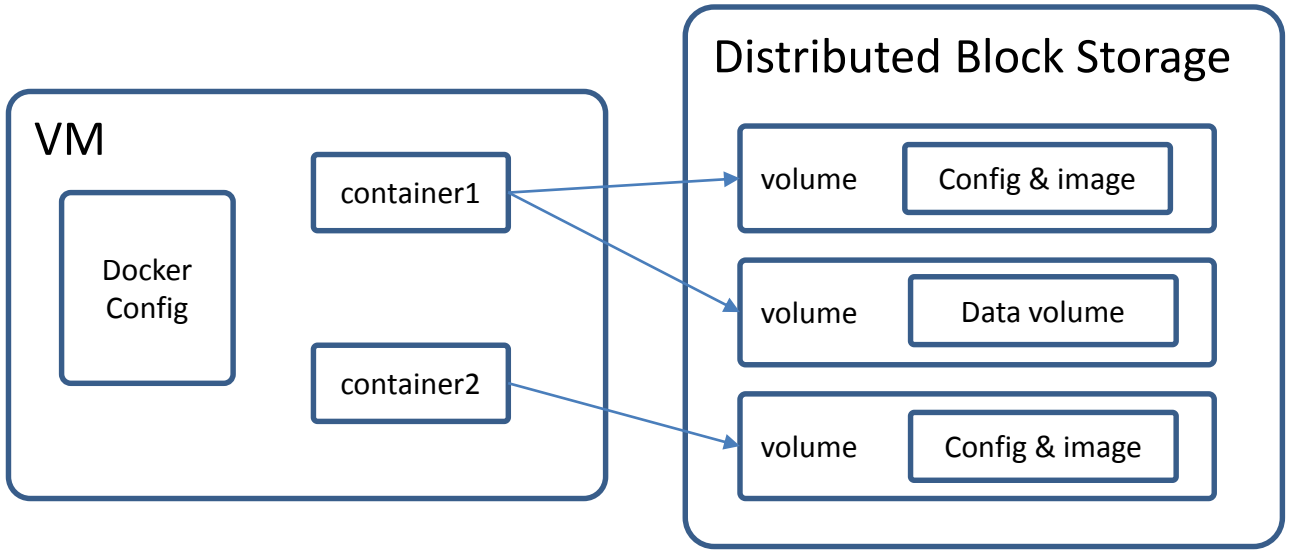
虚拟机启动优化





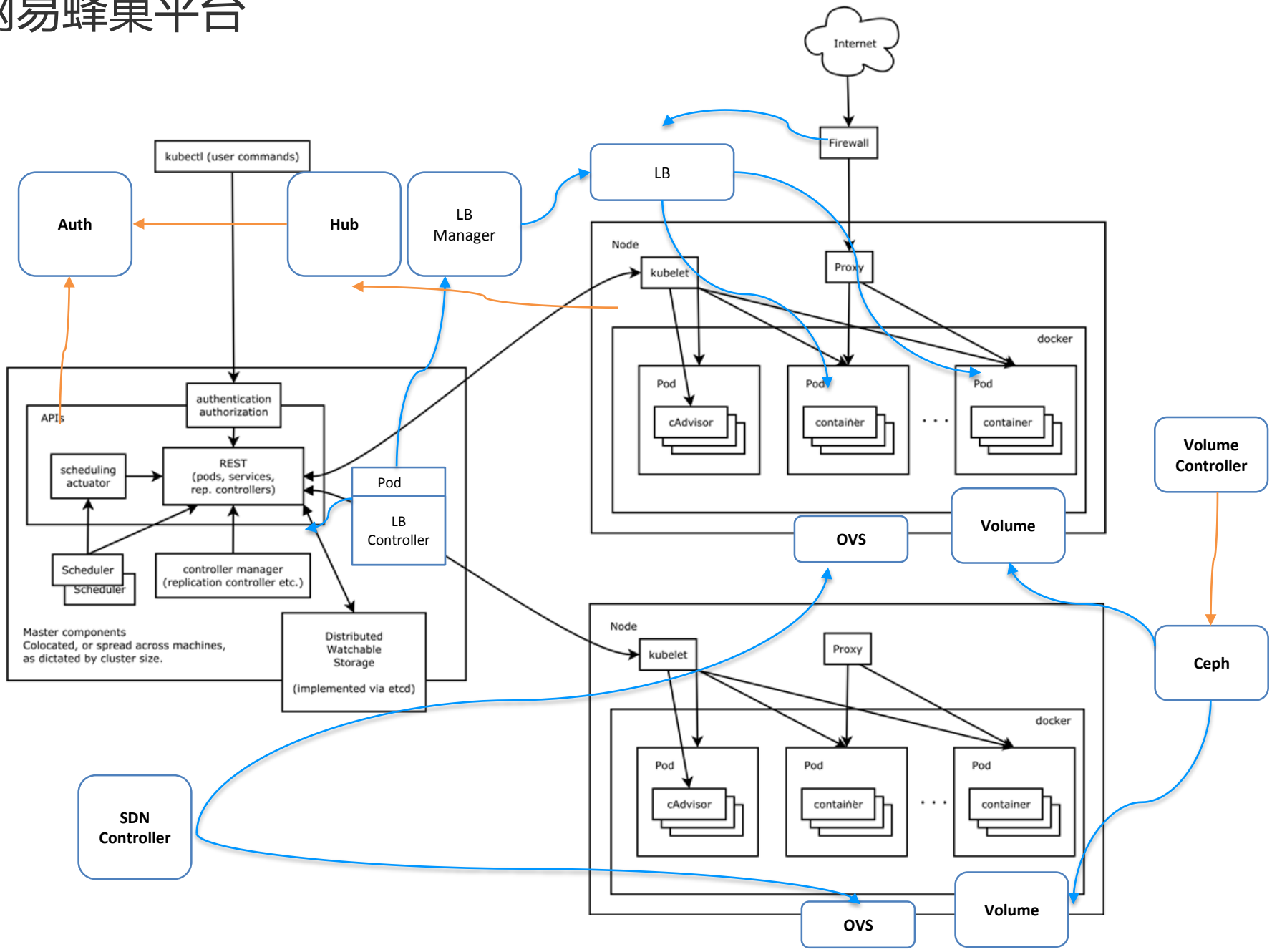
1. 网卡IP初始化
2. 网络路由注入
3. DNS服务IP配置
4. 网卡udev规则

有状态容器

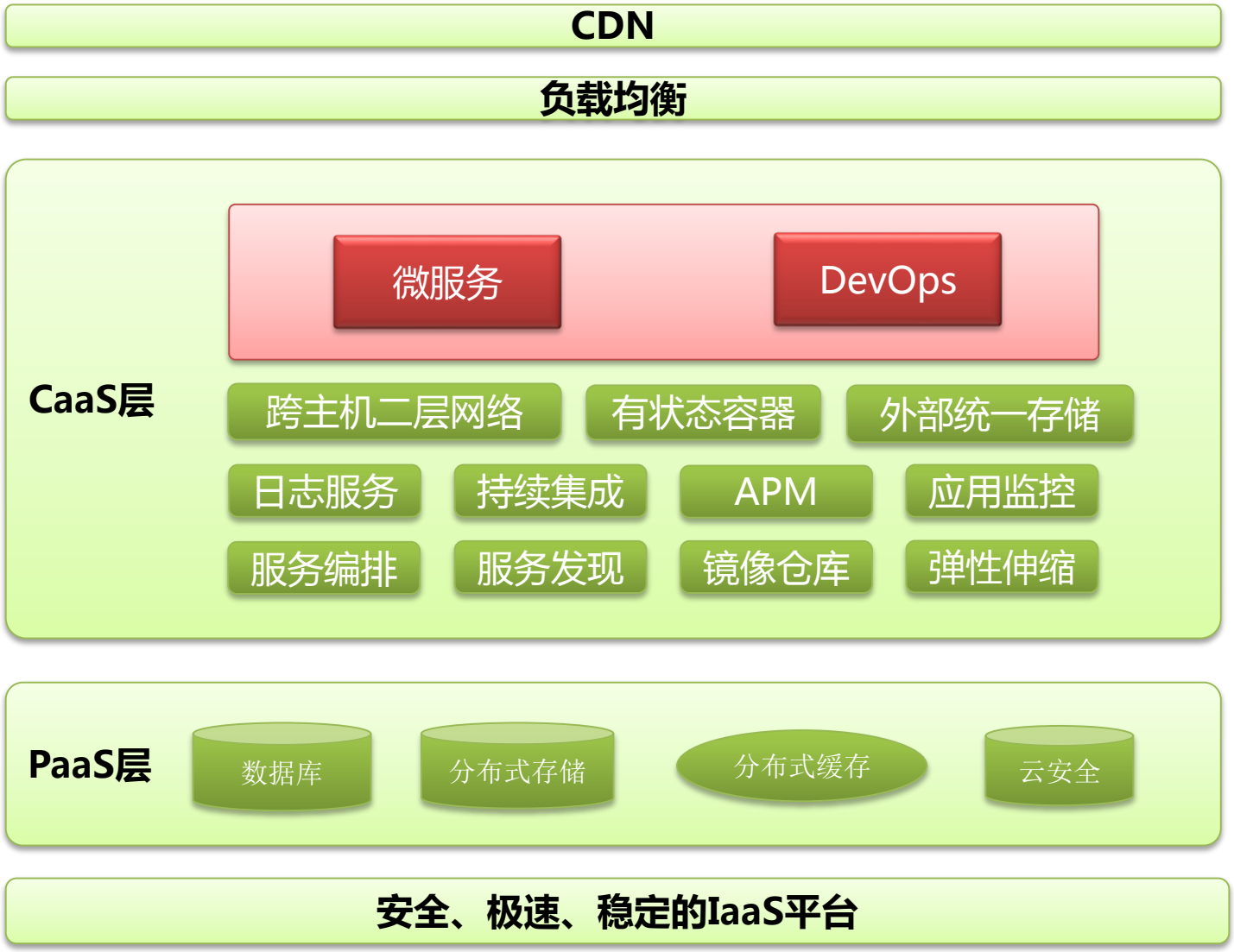


- Node 故障时将远程盘挂载到新 Node，Copy 容器配置信息到 Containers 目录，只有重启 Docker Daemon 才能加载配置启动容器。 解决方法：增加 reload 指令
- Docker daemon 启动时可以通过—graph=指定docker运行时根目录， 但当一个node上运行多个容器时，所有容器的配置信息，文件系统相关数据，数据卷都存放在一个根目录下，导致容器无法独立迁移。解决方法：docker run 增加container-home=dir将容器数据保存在dir目录

网易蜂巢平台


Internet
Firewall
kubectl (user commands)
Auth
Hub
LB Manager
LB
Node
kubelet
Proxy
docker
Pod
cAdvisor
container
Volume Controller
authentication authorization
APIs
scheduling actuator
REST (pods, services, rep. controllers)
Pod
LB Controller
OVS
Volume
Scheduler
controller manager (replication controller etc.)
Master components
Colocated, or spread across machines, as dictated by cluster size.
Distributed Watchable Storage
(implemented via etcd)
Ceph
SDN Controller

蜂巢特色：聚焦应用


CDN
负载均衡
CaaS层
微服务
DevOps
跨主机二层网络
有状态容器
外部统一存储
日志服务
持续集成
APM
应用监控
服务编排
服务发现
镜像仓库
弹性伸缩
PaaS层
数据库
分布式存储
分布式缓存
云安全
安全、极速、稳定的IaaS平台

蜂巢特色：全开源平台

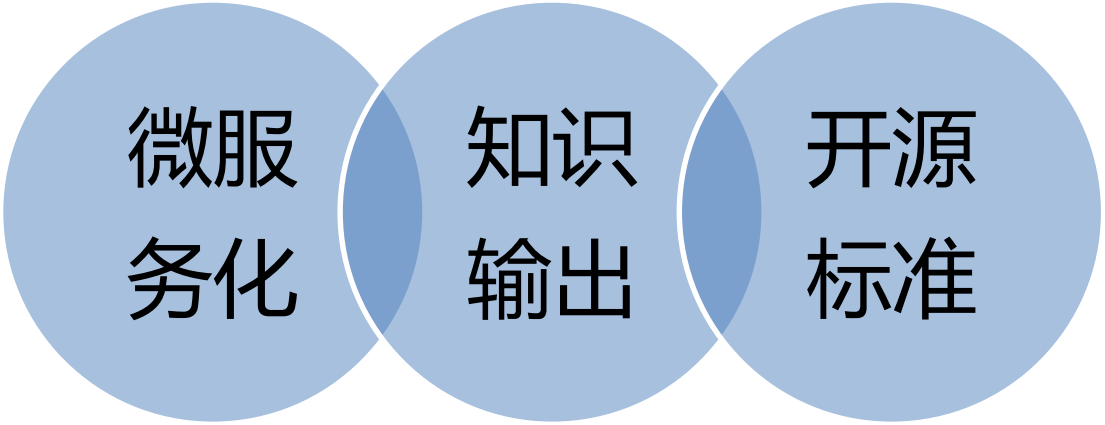
最流行的开源数据库
活跃的社区和日趋完善的功能
网易RDS提供标准接口和稳定保障

最流行的缓存服务
性能最快的Key-Value DB
网易NCR提供标准接口和稳定保障

产品理念完善历史15年
社区热度高Github Star 18200+
企业级应用案例多，最佳实践经验足
透明开源，技术标准化

最流行的开放IaaS平台
100%支持标准API
紧跟社区，新版本发布后3个月平台更新

网易蜂巢助力企业微服务化



PART 04

联系我们

Contact Us

联系我们 Contat Us

C.163.COM

网易蜂巢，欢迎大家使用！

关注“网易蜂巢”微信公共账号 获取网易蜂巢最新动态！

EMAIL: cloudcomb@188.com