网易蜂巢

北京

# 网易蜂巢容器云基础设施优化之道

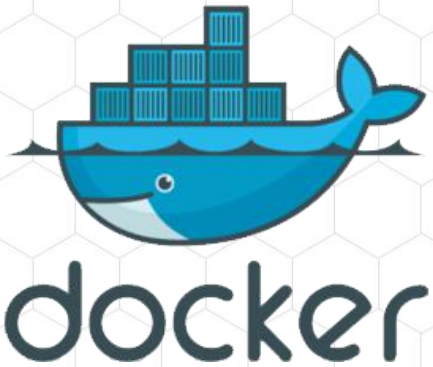
网易蜂巢 / 张晓龙

# 目录

# Docker

## 概念

- 不仅是提高资源利用率、实现资源隔离的轻量级虚拟化，更是一种全新的应用标准化交付方式，是应用交付的“集装箱”




## 现状

- 诞生于2013年，是历史上发展最快的开源软件之一
- 拥有极其活跃的开发者和用户社区
- 形成日益完善的生态系统，获得亚马逊、谷歌等巨头支持

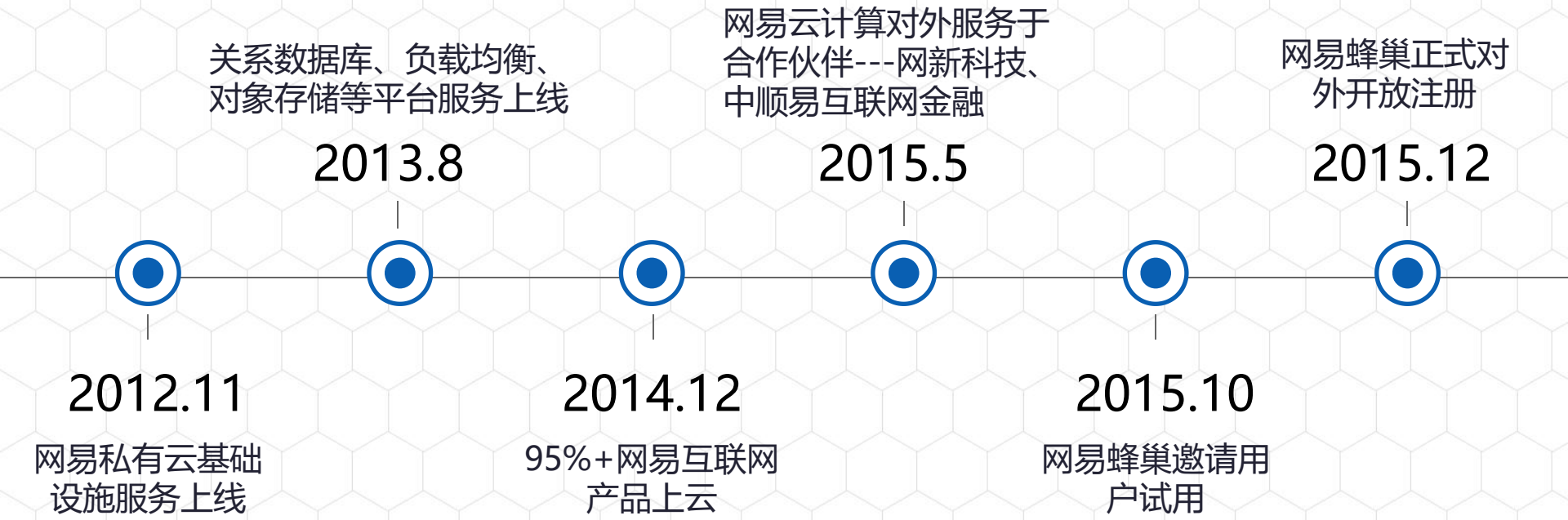
# 网易蜂巢---专业的容器云平台

- 定位
  - 面向高效开发而打造的新型云平台---容器云
- 功能
  - 提供基于容器的应用自动化构建发布、编排管理、镜像仓库等服务
  - 提供业界领先的平台服务如关系数据库、负载均衡、缓存等
  - 提供丰富多样的运维工具如性能监控、报警、日志采集等

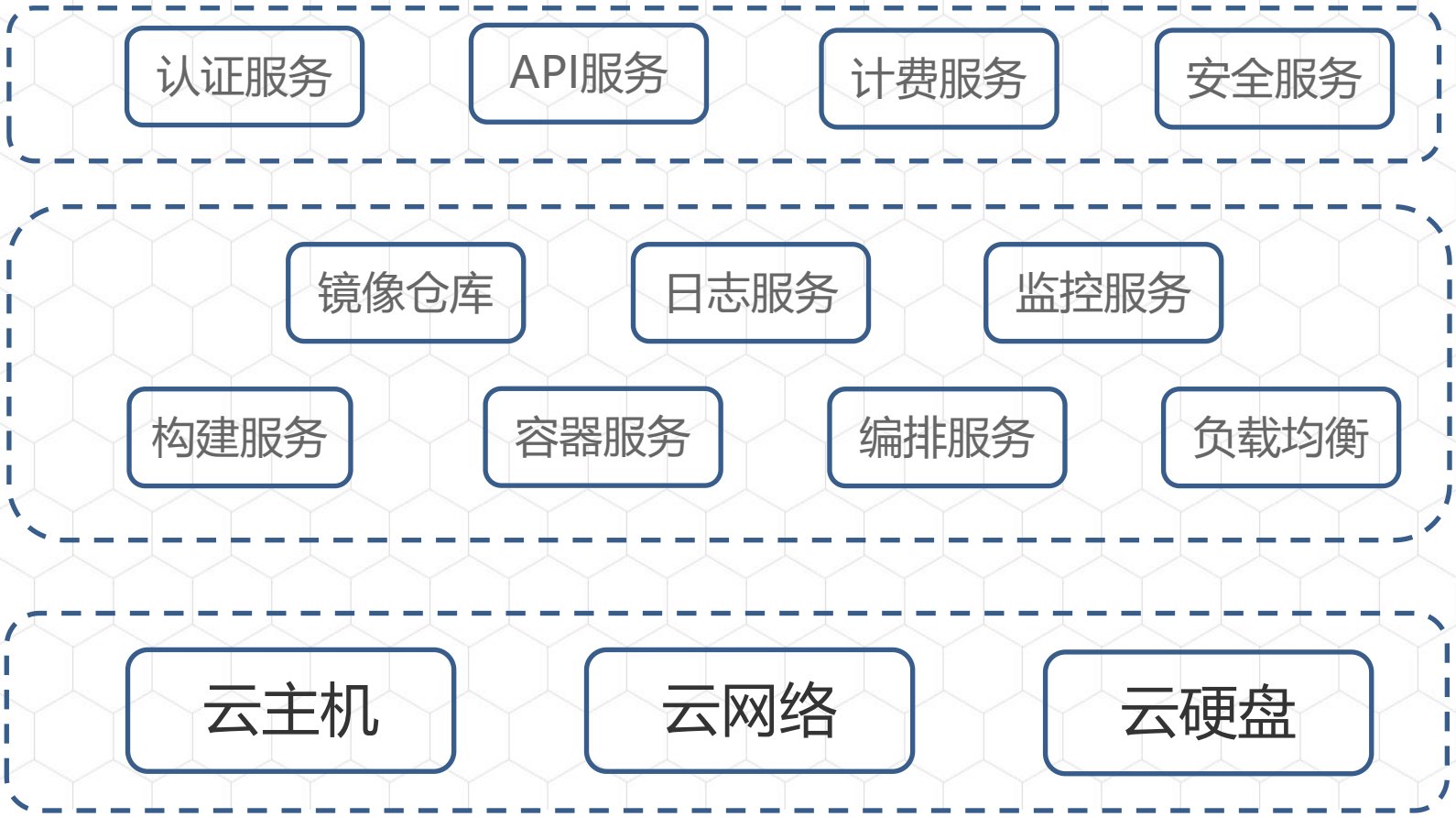
# 蜂巢容器云---优势特色

- 采用Docker容器技术
  - 加速研发全流程，让应用交付更快捷
- 基于网易自研IaaS深度优化
  - 确保极速、安全、稳定的用户体验
- 采用高规格的硬件设备
  - 多线BGP网络接入、万兆网络互联、全SSD存储
- 业界领先的关系数据库
  - 基于网易自研的开源MySQL分支版本InnoSQL深度优化

# 研发历程


2012.11
网易私有云基础设施服务上线
2013.8
关系数据库、负载均衡、对象存储等平台服务上线
2014.12
95%+网易互联网产品上云
2015.5
网易云计算对外服务于合作伙伴---网新科技、中顺易互联网金融
2015.10
网易蜂巢邀请用户试用
2015.12
网易蜂巢正式对外开放注册

# 技术架构


认证服务
API服务
计费服务
安全服务
镜像仓库
日志服务
监控服务
构建服务
容器服务
编排服务
负载均衡
云主机
云网络
云硬盘
服务器
交换机/
路由器
硬盘

# 核心技术

- 基础设施：提供容器运行所需计算/存储/网络资源
  - 虚拟化技术：KVM、OpenVSwtich、Ceph
  - OpenStack
- 容器：资源交付的最小单位
  - Docker
- 容器编排：实现容器集群发布、回滚、迁移、扩容、缩容等
  - Kubernetes






# 容器隔离

- 设计
  - 容器运行于隔离性更强且基于硬件虚拟化技术的云主机
  - 在一个云主机上只运行同一个租户的容器
- 好处
  - 获得更好的容器安全性
  - 故障隔离
  - 可把系统能力如iptable等开放给用户
- 缺点
  - 会带来资源和性能上的损耗

# 容器启动速度优化

- 问题
  - 容器运行于云主机，容器启动依赖于云主机先启动
  - 基于硬件虚拟化技术的云主机启动速度较慢
- 启动速度优化
  - 定制系统镜像，裁剪不必要服务启动加载项
  - 实现云主机IP静态化，加速网络初始化过程
  - 优化OpenStack创建云主机流程
- 效果
  - 运行容器的云主机平均启动耗时在十秒之内

# 容器网络

- 私有网
  - 虚拟扁平二层网络
  - 租户100%隔离
  - IP分配自主控制
- 公网
  - 所有租户共享
- 实现
  - 基础设施云网络服务提供两张网络；
  - 在容器所在云主机创建网络端口并将端口置于容器的网络命名空间

# 容器网络安全

- 网络过滤
  - L2过滤：确保报文源MAC地址是系统所分配端口MAC地址，防止ARP欺骗
  - L3过滤：确保数据包源IP是系统所分配IP，防止IP地址欺骗
  - L4过滤：过滤指定的TCP/UDP端口，便于实施网络封禁
- DDoS攻击防护
  - 基于Intel DPDK实现DDoS攻击防护

# 容器网络宽带QoS

- 网络带宽QoS设计原则
  - 保证用户所申请网络带宽
  - 有效利用空闲网络资源，免费提升用户带宽体验
- 处理网络小包过载
  - 问题：VXLAN小包处理性能不够好，网络小包过多导致宿主机CPU过载（软中断过多），影响网络性能和稳定性
  - 方案：限制容器网络的PPS （Packet Per Second）
- 实现
  - 基于Linux TC并修改OVS，实现保证速率、最大速率
  - 将小包按照MPU（Minimum Packet Unit）大小来处理

# 容器存储性能优化

- 问题
  - Ceph在osd进程重启时会出现长时间、极其严重的性能衰减（80%+）
- 原因
  - osd重启时要恢复重启期间脏数据对象，会消耗大量网络/磁盘开销
- 优化
  - 在pglog记录重启期间数据对象的增量数据，在重启时增量恢复数据对象
- 效果
  - 减少重启过程对集群正常I/O性能影响（ I/O性能降低10%~20%以内）
  - 缩短重启恢复所需时间（重启单个osd从10分钟减少到40秒左右）

# 容器支持运行有状态的服务

- 需求
  - 持久化容器本身数据（运行环境、配置数据等）
  - 快速备份以及恢复容器本身数据
  - 支持有状态容器的迁移
- 方案
  - 定制Docker，将容器数据存储到保证数据可靠、有备份/恢复能力的云硬盘，同时实现有状态容器的迁移
- 实现
  - 增加容器数据目录参数，在容器启动时将数据保存在指定目录中，实现将容器数据存储到云硬盘
  - 增加重载容器配置指令实现无需重启daemon即可加载被迁移容器的配置，解决迁移有状态容器时的重启容器问题

# 容器编排

- 完善多租户支持
  - 实现将node、存储、网络等集群共享资源的租户隔离
  - 完善租户资源的安全访问控制、为每个租户实现独立的认证和授权
- 调度器/控制器并行处理优化
  - 将面向集群的串行调度优化为多租户并行调度
  - 将副本队列串行处理优化为按照多优先级队列并行处理
- 使用多个etcd集群
  - 将Pod/Node/RC等资源拆分到不同的etcd集群存储维护

# 后续计划

- 计算
  - 进一步优化容器启动速度（20s以内）
- 网络
  - 提供更高级网络功能比如虚拟路由器、防火墙等
  - 进一步优化容器网络性能
- 存储
  - 减少Ceph存储集群对物理CPU的高消耗，提升集群整体I/O性能
  - 实现Ceph数据重分布的流控，使集群扩容时I/O性能更平滑
- 容器
  - 实现Docker版本热更新
- 容器编排
  - 进一步优化容器编排性能以及提高可扩展性

网易蜂巢

谢谢观看！

扫一扫，关注我们