网易蜂巢：构建公有容器云实践

张晓龙

目录

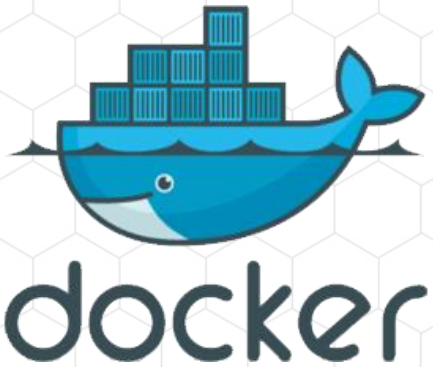
容器

容器

- 容器是提高资源利用率、实现资源隔离的轻量级虚拟化
- 容器封装了完整的应用运行环境（操作系统、库、运行时、业务代码），是应用交付的“集装箱”
- 容器改变了应用的管理和部署方式

Docker

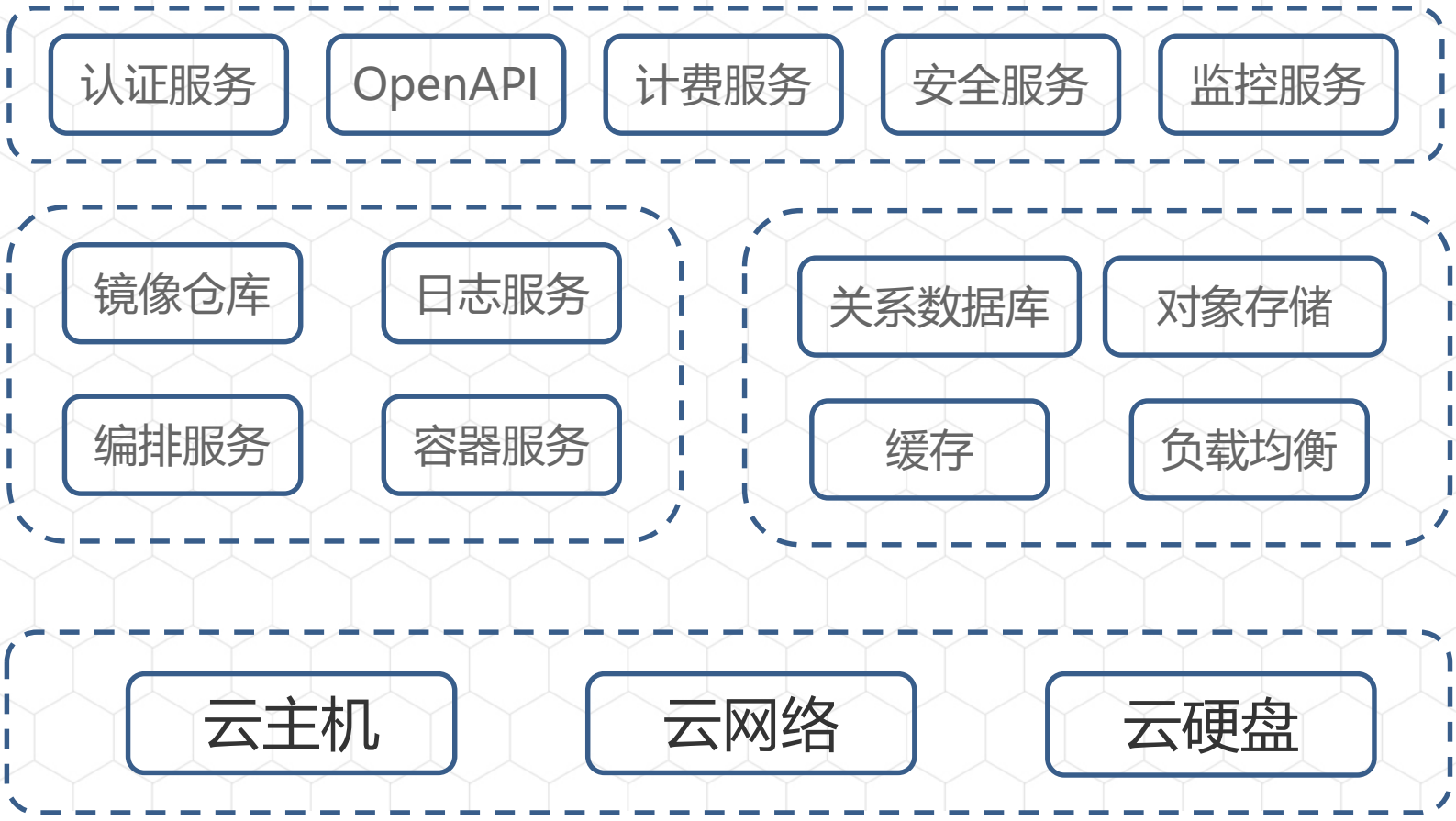
- 诞生于2013年，是历史上发展最快的开源软件之一
- 拥有极其活跃的开发者和用户社区
- 形成日益完善的生态系统，获得亚马逊、微软、谷歌等巨头支持



网易蜂巢---容器云

- 定位
  - 面向高效研发而打造的新一代云计算平台，提供弹性计算、DevOps工具链及微服务基础设施等，帮助企业解决IT、架构、运维等问题，使企业更聚焦其业务
- 功能
  - 提供容器及其镜像加速、镜像构建、镜像仓库等在内的容器服务
  - 提供包括对象存储、CDN、关系数据库、负载均衡、缓存服务、安全服务等在内的完善平台服务
  - 提供包括服务发现、编排服务、APM服务、持续集成、监控服务、日志服务、持续发布等在内的完整DevOps工具链

服务组件


认证服务
OpenAPI
计费服务
安全服务
监控服务
镜像仓库
日志服务
编排服务
容器服务
关系数据库
对象存储
缓存
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

- 容器：平台资源分割/交付的最小单位
  - Docker
- 容器编排：实现容器集群管理的/扩缩容/灰度升级/服务发现/故障恢复等功能
  - Kubernetes
- 基础设施：提供容器运行所需计算/存储/网络资源
  - 高效管理资源，确保资源按需分配、弹性交付
  - 确保交付资源的服务质量（QoS），如计算能力、网络性能、I/O能力等
  - 采用运行于通用硬件设备的软件定义技术：OpenStack、KVM、OpenVSwtich、Ceph

关键技术-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

关键技术-容器网络

- 私有网
  - 特点：虚拟扁平二层网络、租户100%隔离
  - 实现：node上挂载私有网卡并建立网桥，网桥上加veth pair实现（一端在容器、另一端在网桥上）
- 公网
  - 特点：所有租户共享
  - 实现：将云网络的外网端口放到容器namespace

关键技术-容器存储

- 需求
  - 提供持久化容器数据的能力
  - 支持有状态容器的迁移
- 方案
  - 实现指定rootfs目录启动Docker容器，将云硬盘挂载点设置为指定的rootfs目录，借助云硬盘的备份能力实现容器数据备份和恢复
  - 解决容器迁移时在迁移目标节点需要重启Docker Daemon的问题

关键技术-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技术实现高性能实时抗攻击

关键技术-容器网络带宽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
- 实现
  - 基于Linux Traffic Control 并修改OVS，实现保证速率、最大速率
- 处理网络小包过载问题
  - 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）

关键技术-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创建主机：解决节点定时任务执行过长导致创建请求被延迟处理的问题
- 效果
  - 运行容器的云主机平均启动耗时在十秒之内

关键技术-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）

关键技术-容器编排优化

- 完善多租户支持
  - 实现将节点、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
- API Server 优化
  - 开启API Cache以及使用新版本支持多API Server的特性

谢谢观看！