中国移动
China Mobile

中国移动在云原生操作系统的探索与实践

演讲人

任林 信息技术中心 高级研发工程师

刘志磊 在线营销服务中心 高级研发工程师

www.10086.cn

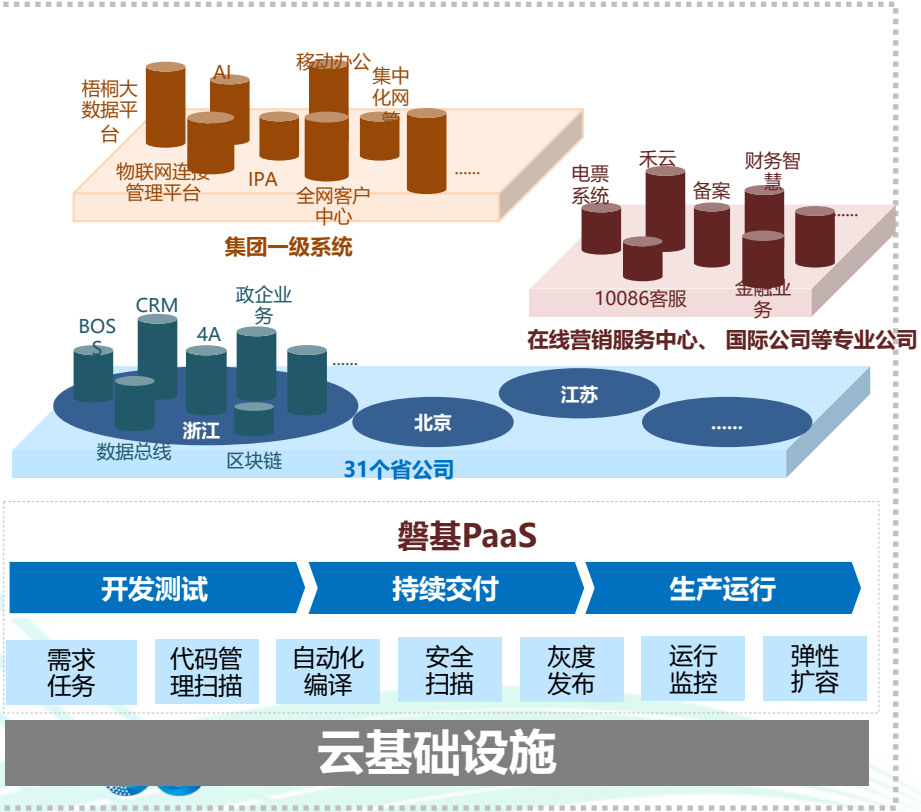
中国移动
China Mobile

目录 CONTENT

背景——中国移动磐基PaaS平台

中国移动 China Mobile

磐基平台是中国移动集团10个战略级重点项目之一，平台持续演进探索算力网络等前沿技术领域，支持单应用系统日峰值交易量达3亿笔，月总交易量达60亿笔，支撑规模和并发能力均达到业界领先水平。



规模业界领先

资源池数	节点规模	覆盖B/O/M/D系统	K8S集群	容器实例数	管理代码
14个	2.62万个	407个	387个	38万+个	10亿+行

能力成熟稳定

多集群管理　应用统一部署编排　弹性扩缩容　组件管理　微服务管理　运维监控

已获得云原生社区CNCF一致性认证、信通院容器云可信云认证、DevOps先进级认证等多项行业认证与奖励

价值成效突显

赋能一线生产	50% 业务上线效率（配置类）	15% 业务上线效率（新业务）	客户无感知 应用发布	秒级 资源配备	分钟级 容灾切换
IT系统治理	30% 利用率提升	40% 技术组件收敛	全网统一 技术标准	组织管理 研发资产	自主可控 技术能力

背景——痛点

租户应用、组件的生命周期，得以通过我们的PaaS平台统一进行管理。然而，对于PaaS集群的承载者，K8S集群节点的操作系统，却仍在采用着传统的分散式管理模式。

规模压力

随着磐基平台向域内域外推广速度的加快，所纳管的K8S集群节点也 成倍增加，如何有效管理这些节点的操作系统版本，降低运维工作量?

安全驱动

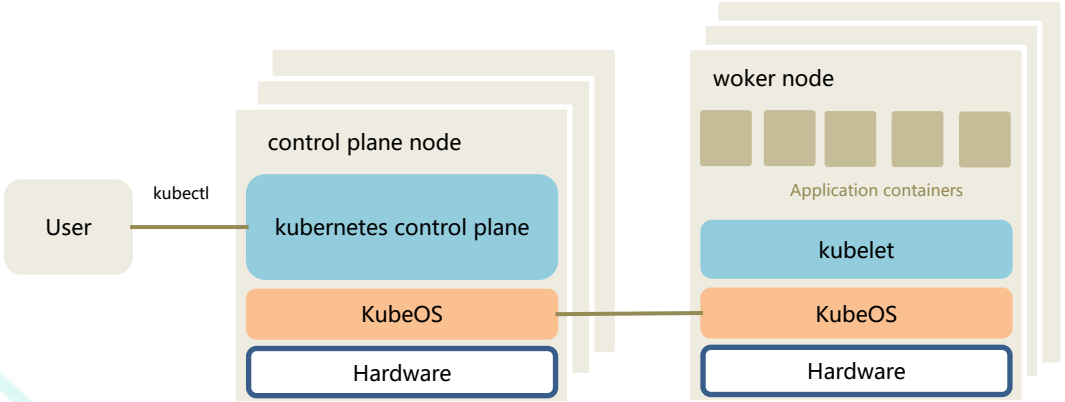
全网不定期和周期性的安全基线扫描发现的安全漏洞，补丁修复，涉及了所有的集群和主机节点。对传统的主机管理模式提出了更大的挑战。

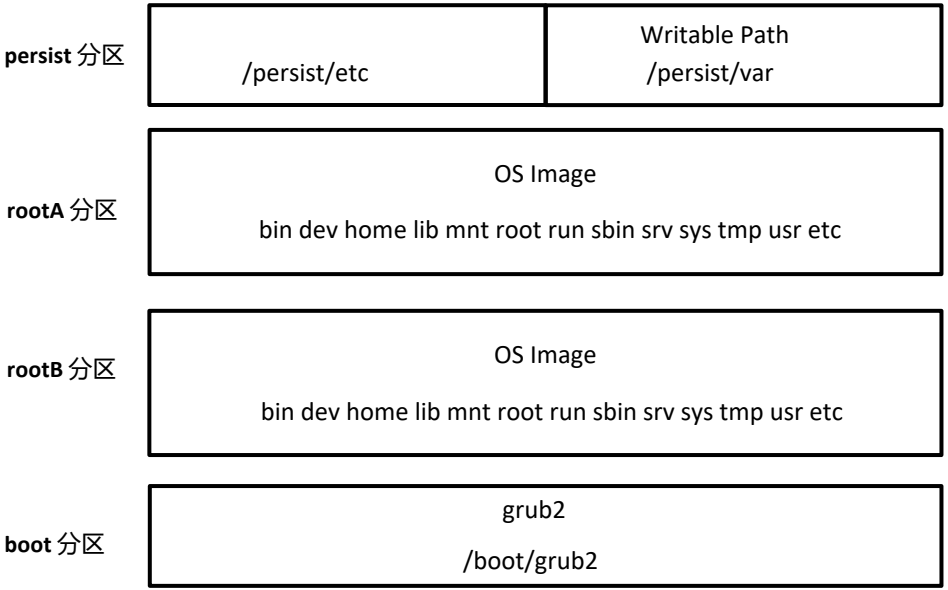
瘦身需求

作为K8S节点宿主机的操作系统，传统的安装模式会带来很多冗余安装包，增加安全风险和漏洞修复工作量。

背景——KubeOS：基于openEuler的容器化操作系统

- KubeOS是OpenEuler下的一个开源项目，他将宿主机的OS作为Kubernetes一个组件，通过Kubernetes，统一管理容器和宿主机OS，使宿主机OS和容器业务调度协同工作，实现了宿主机OS版本的原子化管理，提升了对整个集群的管理效率。
- KubeOS采用了OS的原子升级策略，即，不提供包管理器， 软件包变化即OS版本变化。
- OS轻量化，减少不必要组件，快速升级，替换。





探索——ETC目录困境

在KubeOS的设计中，etc目录需要与/bin,/usr等只读分区目录分开，单独挂载成可写的，这样本地化配置才能固化下来。但是etc目录的挂载时机却成为一个难题。

Moving /etc to separate partition

Asked 10 years, 6 months ago Modified 5 years, 8 months ago Viewed 8k times

4

How to put `/etc` on separate partition? Obviously i can't do that by editing `/etc/fstab` like i did with `/home`, because... it's in `/etc`. I want `/etc` and `/home` on one partition (sda7), and the rest on the other (sda6). I guess `/etc` must be symlink to `/mnt/part2/etc` (`/mnt/part2` being mount point of sda7), and same with `/home`. But how to tell the system to mount part2 without access to `fstab`?

I'm using Arch Linux x64, if that helps.

partition fstab etc

7

> I want /etc and /home on one partition

No you don't. It's like asking to have your brain transplanted to your knee :-) Whatever your problem is, making `/etc` a separate partition or merging it with `/home` is not the solution. What is the *actual problem you want to solve*?

Share Improve this answer Follow

answered May 30, 2013 at 21:57

Jens

1,758 ● 4 ● 17 ● 36

- 用户在使用虚机过程中，可能需要需要额外空间，另外，一些K8S的特殊功能节点，例如Ceph节点或数据库节点，为集群应用提供网络存储或数据库能力，需要单独挂盘。
- systemd方式 vs fstab方式。理论上，两种方式都是可以支持的。
- KubeOS系统分区都是通过Systemd进行挂载。
- 所有的挂盘方式都与etc目录有关。

探索——从KubeOS到云原生OS管理系统

中国移动 China Mobile

KubeOS除了实现了Operator化管理之外，还提供了一套脚本化制作OpenEuler镜像的框架。但用于生产却不够。

镜像制作优化

多操作系统支持

- 从OpenEuelr到CentOS、BCEuler，Kylin，支持不同的操作系统：
- 不同操作系统的GRUB2 UEFI引导目录不尽相同，需要在镜像制作过程中和os-agent代码中，针对不同的操作系统类型进行适配。

OS镜像分区优化

- KubeOS缺省的镜像只有20G，但实际生产系统中用到的虚机操作系统一般是100-200G不等：
- 扩大ROOT-A和ROOT-B分区。
- 根据生产实际，增加对分区格式的支持，支持xfs文件系统。

镜像制作流程优化

- 将一个大的编译流程文件分成若干阶段，增加灵活性
- 支持非yum安装模式。
- 在镜像编译过程中增加了自定义脚本，支持自定义步骤。

OS管理功能

升级回退功能增强

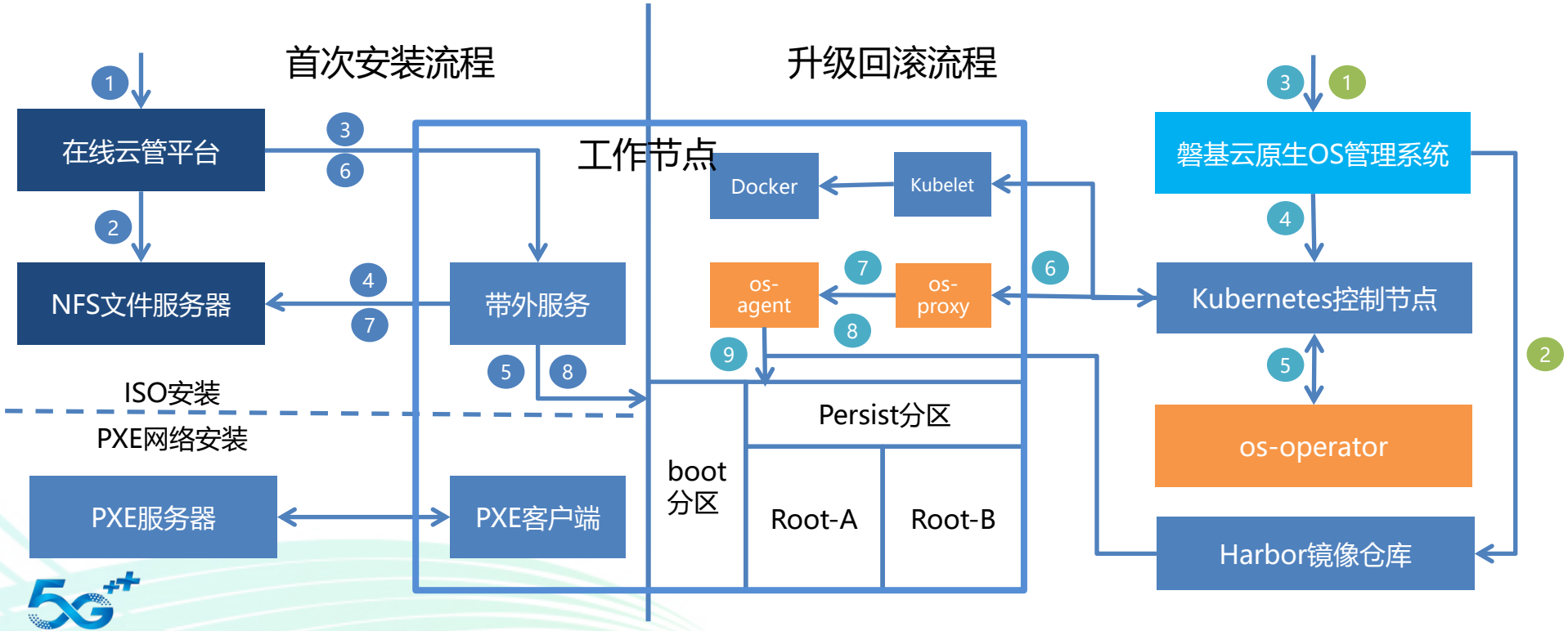
- 与社区一起讨论，将最初的全集群一次性升级或回退逻辑，优化为可部分升级或回退。
- 可以通过节点标签方式指定本次要升级/回退的节点。

集群节点分类管理

- 在实际操作中，有些集群节点的操作系统较为特殊，切无法KubeOS化管理，需要单独处理。。
- 在管理系统上增加了筛选逻辑，可以将指定节点从集群列表中隐藏，保证了这些节点OS的独立性。

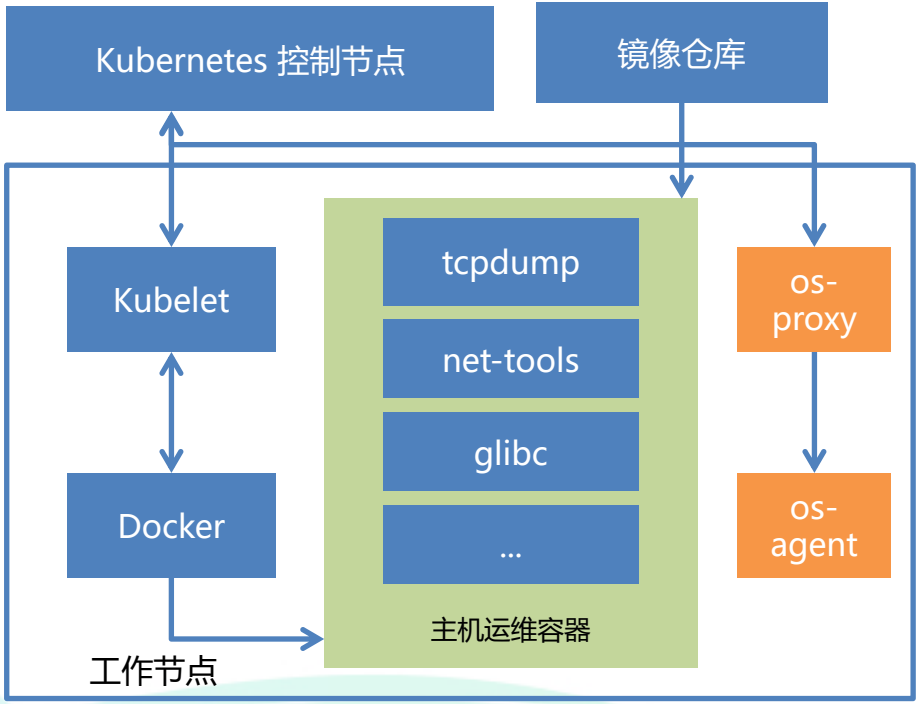
亮点——裸金属的实践思路

结合在线容器基础设施环境，进行基于ISO镜像的首次装机尝试，打通裸金属服务器的“自助装机+自动升级”操作系统管理流程，提升基础设施自动化水平。



亮点——问题排查方式

云原生OS管理系统为了降低系统运行的安全风险，大幅精简容器主机系统软件安装，为提升主机问题排查手段，使用主机运维容器形式进行日常故障排查分析。



- 实现原理
  主机有问题排查类需求时，通过在主机上运行共享Namespace的运维容器，使用运维容器用运维容器内命令排查主机问题。
- 下一步计划
  使用sysmaster+KubeOS的admin-container容器方式替代现有解决方案，以KubeOS形式提升问题排查方式及问题排查效率

展望

基于KubeOS构建的磐基云原生OS管理系统能够较好的解决大规模容器集群主机操作系统管理运维问题，未来将持续协同，做好云原生OS管理系统技术演进。

节点分组

集群内主机可按组划分升级单元，支持同集群内不同操作系统以及不同拓扑域主机独立升级

升级策略

节点分组基础上提供滚动或者灰度升级策略，分批完成同组内节点升级，降低节点升级故障影响范围

跨版本升级

支持同一发行版不同版本操作系统升级操作，无缝跟进社区LTS版本系统演进计划

跨系统升级

在双ROOT分区支持原子化升级技术基础上，探索不同发行版操作系统原子升级可行性，提升操作系统替换便捷性