与「十」俱进，码出未来

2018源创会年终盛典

主办方 开源中国 oschina.net

腾讯 Kubernetes从开源到落地

腾讯　张超
dockerzhang@tencent.com

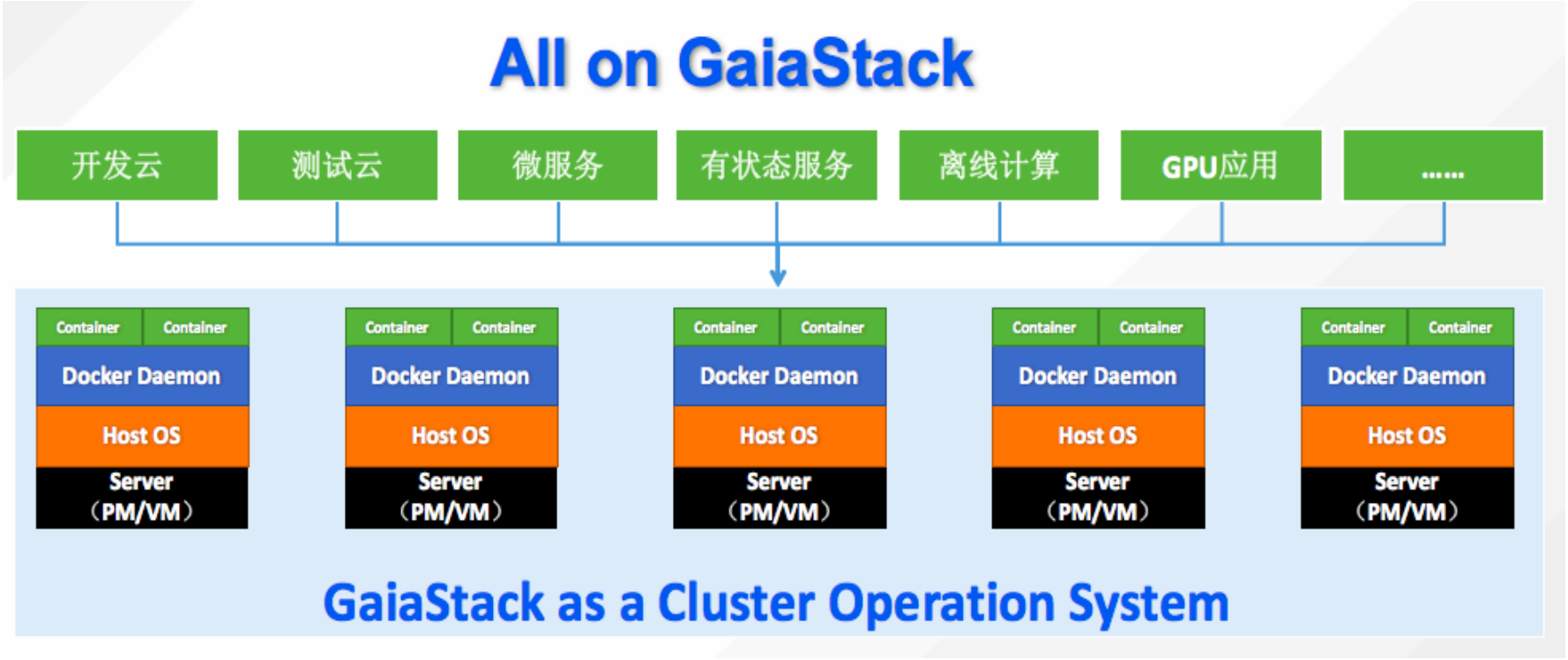
内容

1. GaiaStack平台简介

2.社区版 k8s在腾讯的演进

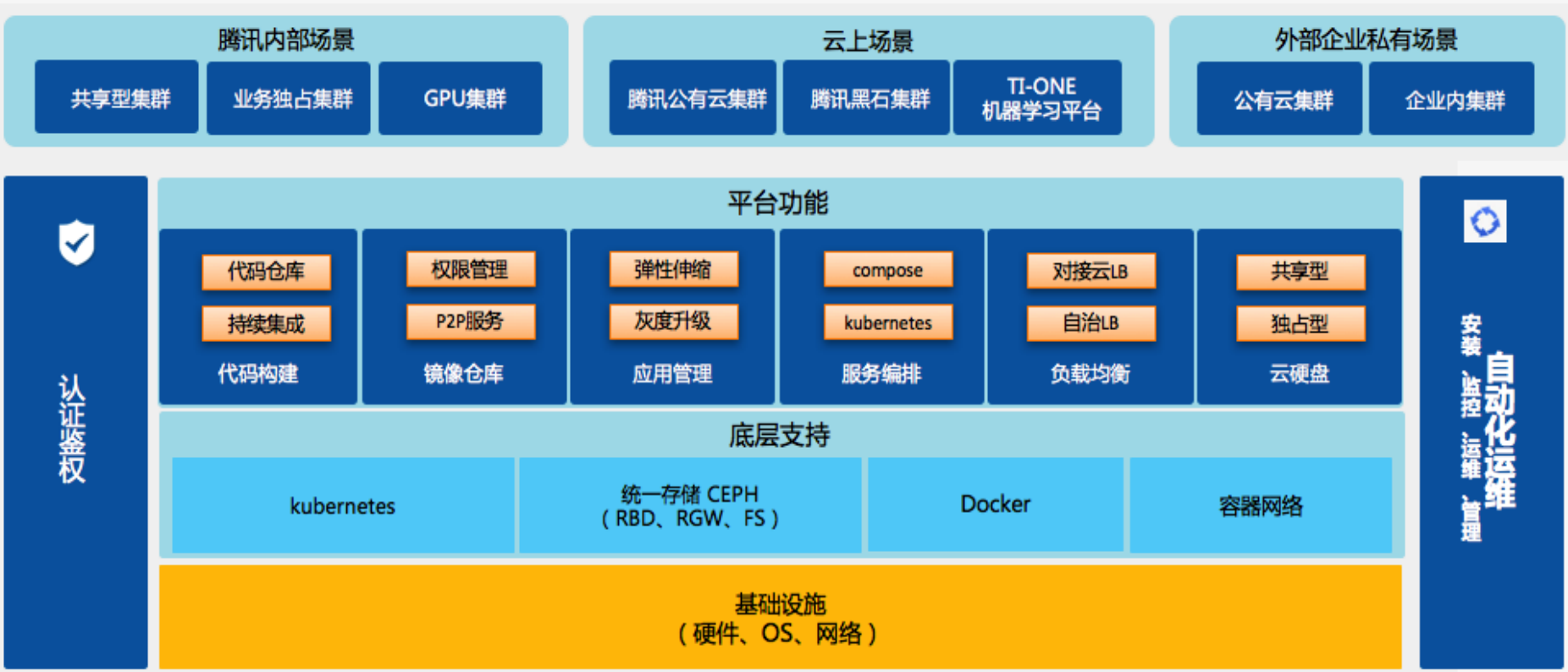
3. GaiaStack应用案例

GaiaStack平台简介



Gaia是腾讯基于Hadoop Yarn为大数据平台提供资源的一套底层资源调度系统。
GaiaStack是腾讯基于kubernetes打造的“支持多场景，适应全应用”的新一代企业级容器云平台。

GaiaStack平台简介



社区版Kubernetes私有云场景挑战

- 资源纬度仅支持CPU、Memory、ephemeral-storage（1.12 beta），网络出入带宽和磁盘IO如何管理？
  - ✓http://2017.qconbeijing.com/presentation/512
- Deployment、Statefulsets、Job、CronJob应用类型真的好用吗？
- Flannel、calico、weave容器网络选用哪个？负载均衡用什么？
- 云硬盘用哪个？
- 升级版本不兼容？
- 日志，监控，告警问题

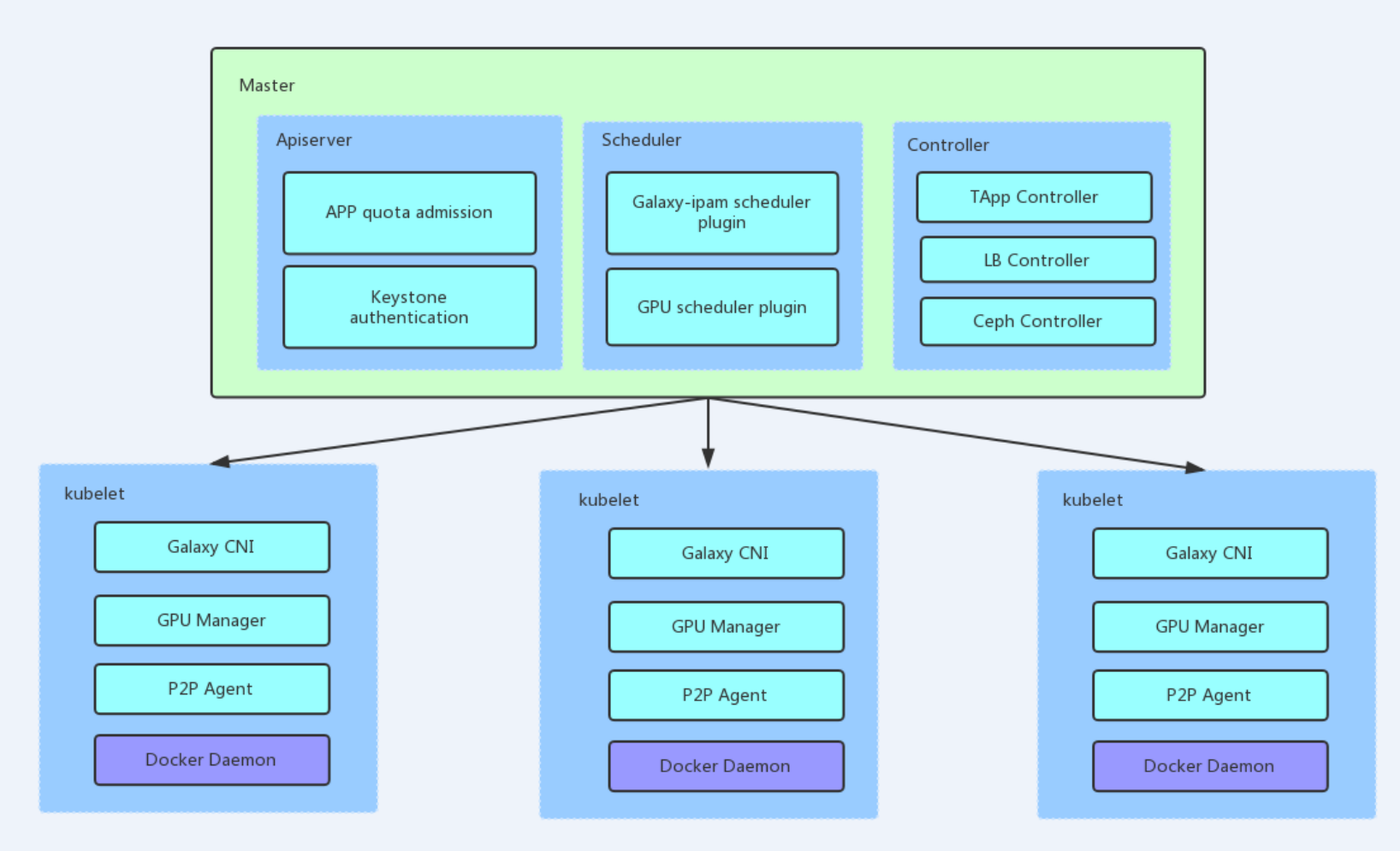
内容

1. GaiaStack平台简介

2.社区版 k8s在腾讯的演进

3. GaiaStack应用案例

社区版 k8s在腾讯的演进— GaiaStack k8s VS 社区版 k8s



向社区贡献**40+** PR

社区版 k8s在腾讯的演进—TAPP应用类型

利用Kubernetes CRD功能自研TApp应用类型，与Kubernetes Statefulset应用类型相比：

- **相同点：**

✓ Pod具有唯一自增ID

✓ 绑定单独云盘，迁移时数据盘跟随迁移

- 不同点：

✓ 可以指定实例id做删除、停止、重启等操作

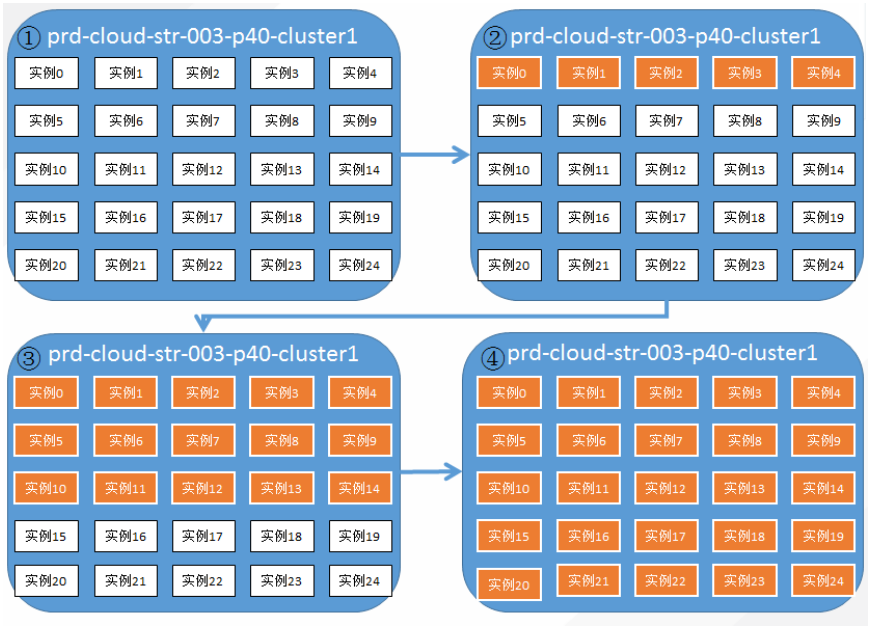
✓ **支持指定若干实例多次进行删除、停止、重启、原地灰度升级、回退Pod等操作**

✓ 单个TAPP应用的Pod支持N个版本

✓ 支持多个版本容器镜像

✓ 单个实例全生命周期跟踪

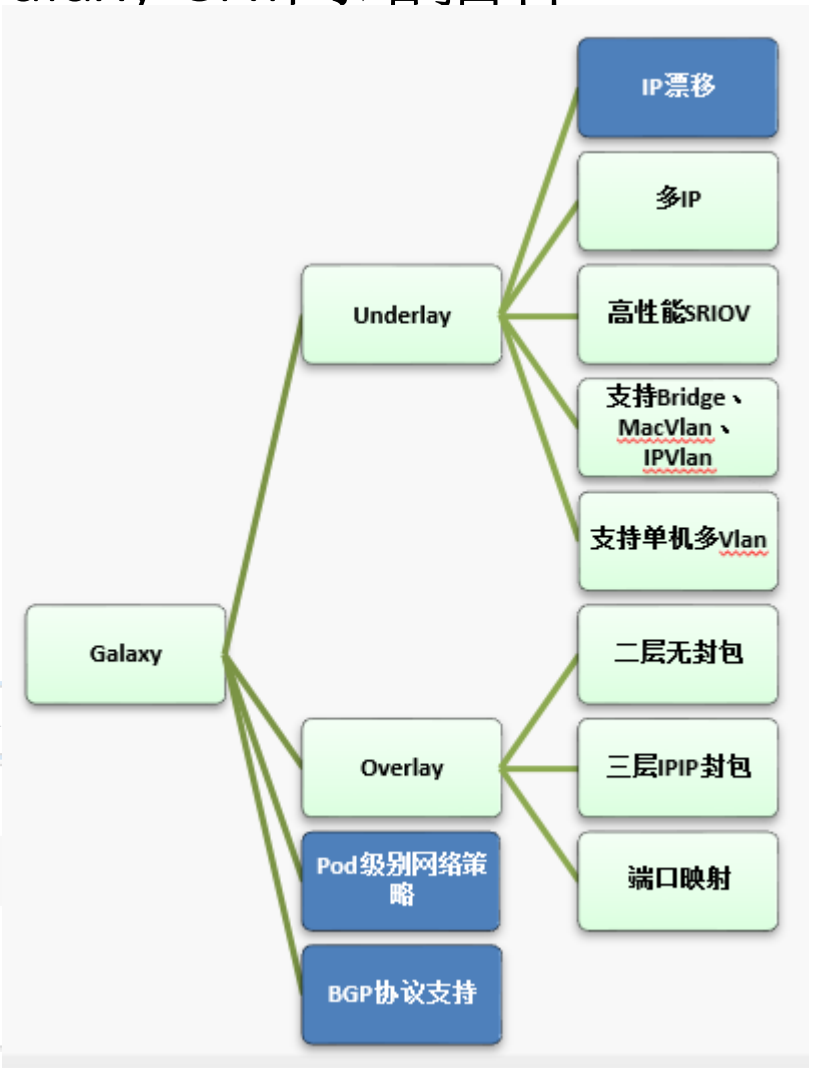
✓ 实现真正的灰度升级/回退



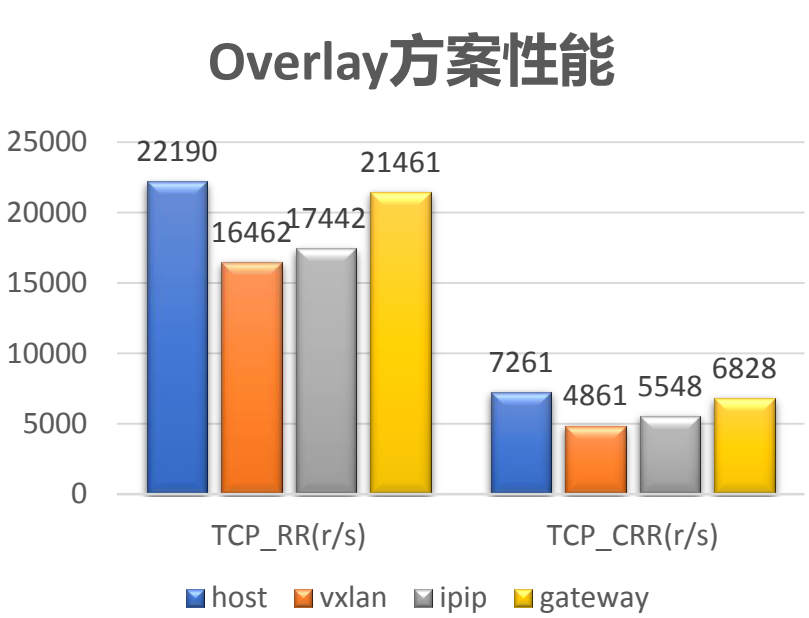
社区版 k8s在腾讯的演进— Galaxy CNI网络插件

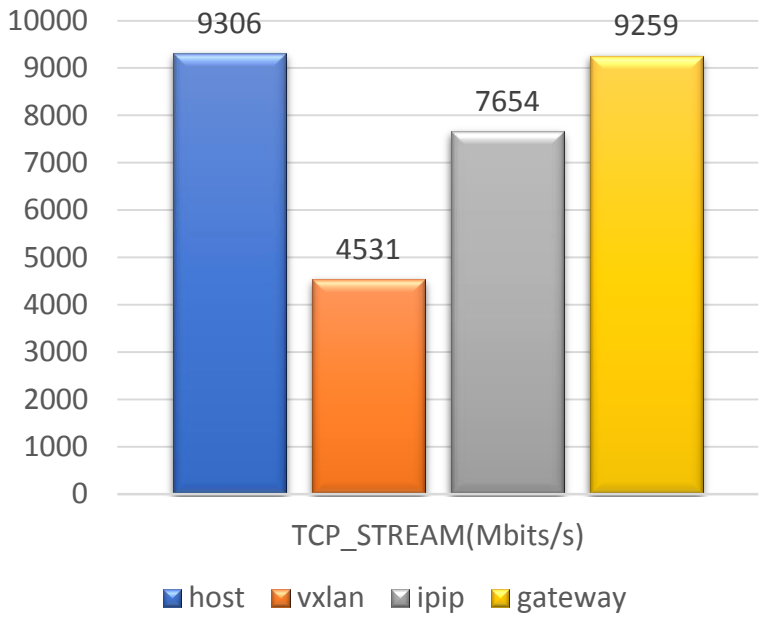
- **因为各种不同场景的需要，自研网络项目Galaxy**
- ✓ 同时提供Underlay + Overlay方案
- ✓ 普适性，多种网络适应不同场景
- ✓ 性能领先

- 选择应用适合的网络方案
- ✓ 不同的应用可以选择不同的网络模式
- ✓ 同一主机的不同容器可以选择不同的网络模式



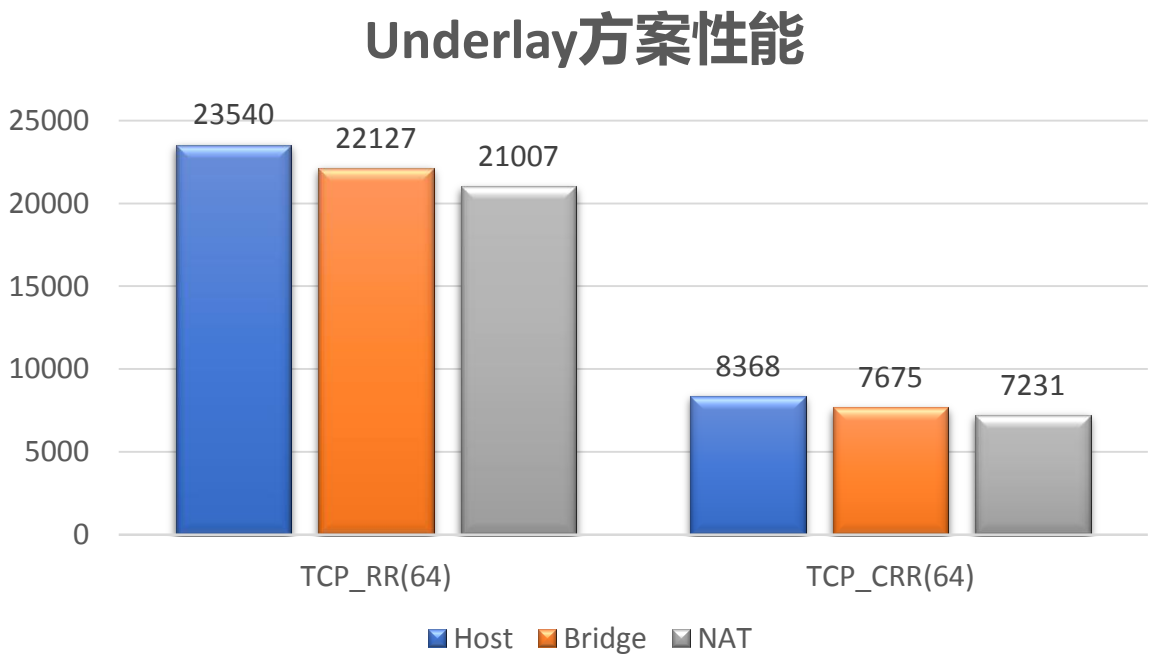
社区版 k8s在腾讯的演进— Galaxy CNI网络插件





GaiaStack提供Overlay方案是IPIP + Host Gateway 混合方案，性能仅次于calico3层方案
短链接 Vxlan比HOST差33%，IPIP比HOST差23%，Gateway比HOST只差6%
方案被社区合并 https://github.com/coreos/flannel/pull/842

社区版 k8s在腾讯的演进— Galaxy CNI网络插件

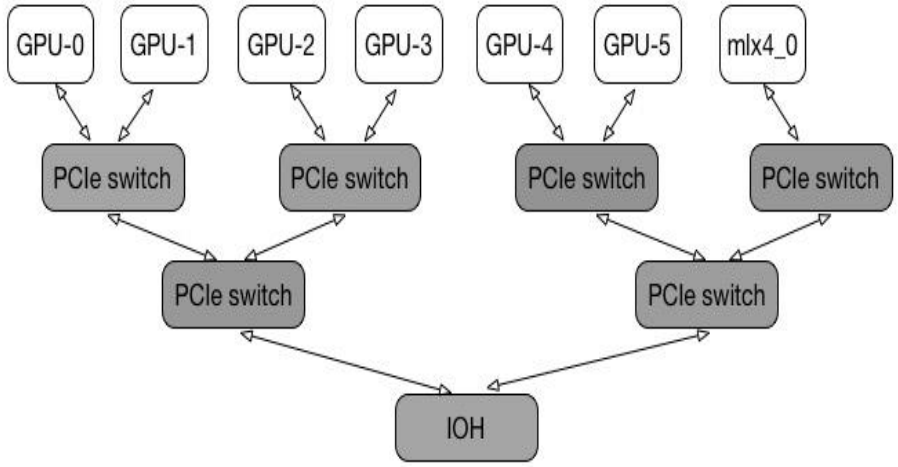


Bridge比Host仅差6%

社区版 k8s在腾讯的演进—GPU Manager设备插件

自研Kube-scheduler调度器插件 + Device Plugin GPU Manager

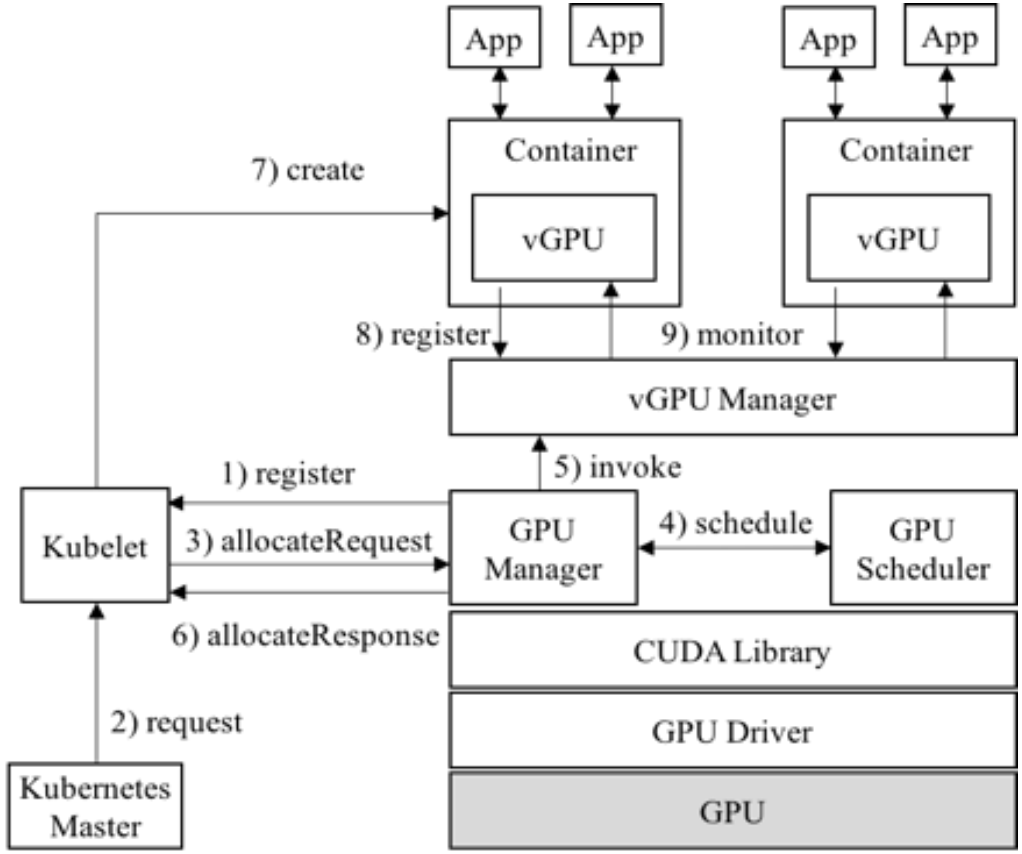
- GPU 拓扑感知调度器
- 异构集群的精细化管理
  - ✓ 管理不同型号的GPU（如：M40、P40等）组成的集群限定不同业务可以使用的卡种类数量
  - ✓ 保证业务能够用到预期的GPU资源（GPU卡数，GPU型号）
- 调度分配优化
  - ✓ Drain node调度解决资源碎片问题
  - ✓ 多卡需求按GPU架构优先分配距离较劲的卡



社区版 k8s在腾讯的演进— GPU虚拟化

- **业内首创容器场景的GPU虚拟化技术**
- 多个程序共享同一张GPU卡，提高资源利用率
- 支持GPU卡，GPU内存的虚拟
  tencent.com/vcuda-core=0-100、
  tencent.com/vcuda-memory=0-100
- GPU卡和内存资源申请通过Manager进行限速，
  用户程序的数据流不经过Manager
- 对用户程序零入侵
- 基于kubernetes device plugin实现，对
  kubernetes、kernel无入侵

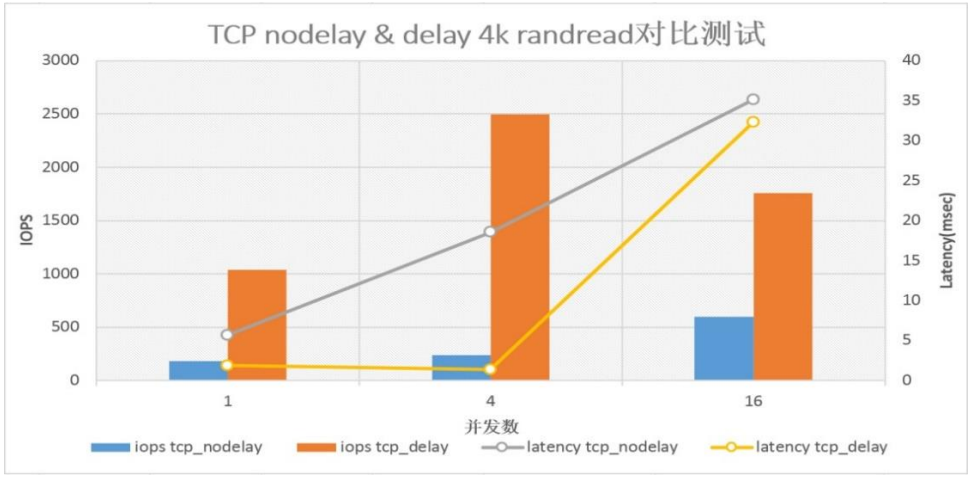
即将在The IEEE ISPA 2018发表的论文《GaiaGPU: Sharing GPUs inContainer Clouds》

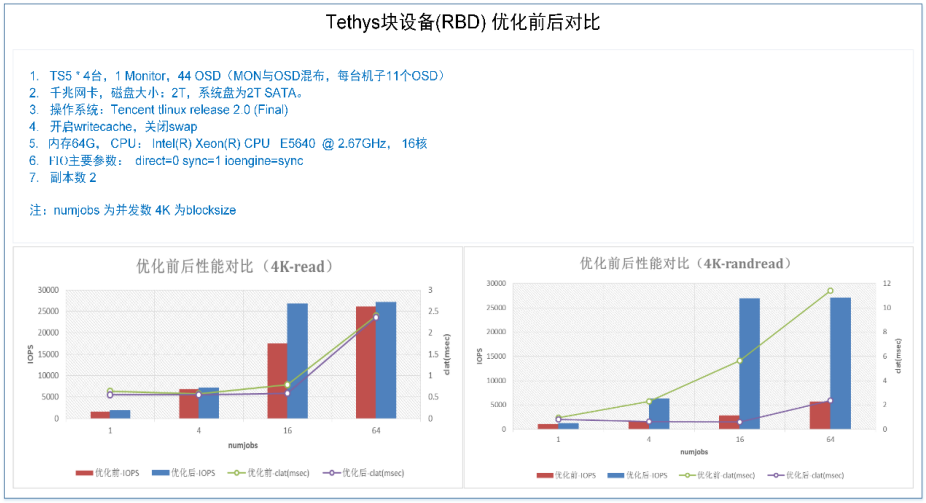


社区版 k8s在腾讯的演进—云盘管理

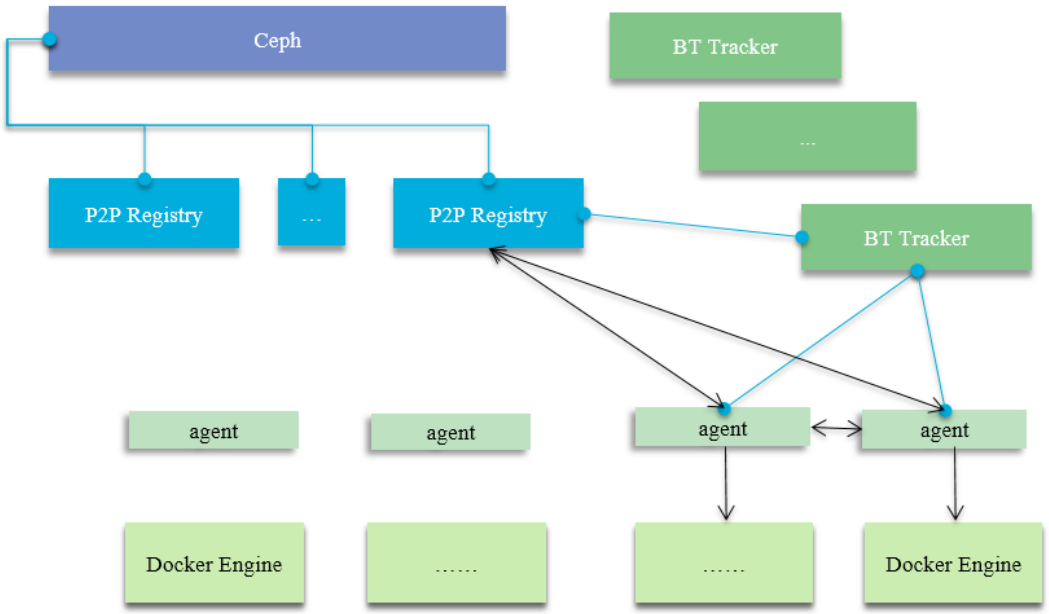
- 支持CephRBD、CephFS，腾讯公有云CBS、CFS等
- 外部controller支持CephFS dynamic provisioning
- Kubelet支持Volume在线扩缩容，https://github.com/kubernetes/kubernetes/pull/62460

- 一名Ceph官方组织成员
- ✓ 提升mds对大量文件的目录处理速度6~10倍
- ✓ 提高了mds主备切换的速度
- ✓ Cephfs内核模块稳定性改进，bug fix
- ✓ 支持quota
- ✓ 支持Jewel，支持keyring挂载权限





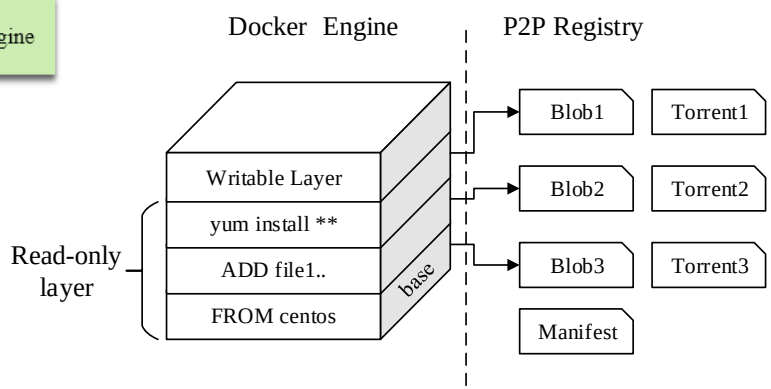
社区版 k8s在腾讯的演进— P2P Docker registry



主要设计思想：

- ✓ 在镜像下载过程中，引入BT协议
- ✓ 在Blob上传时，对Blob生成种子
- ✓ 每层分别做种
- ✓ 在下载镜像的Blob时，先下载种子，再通过种子文件下载数据

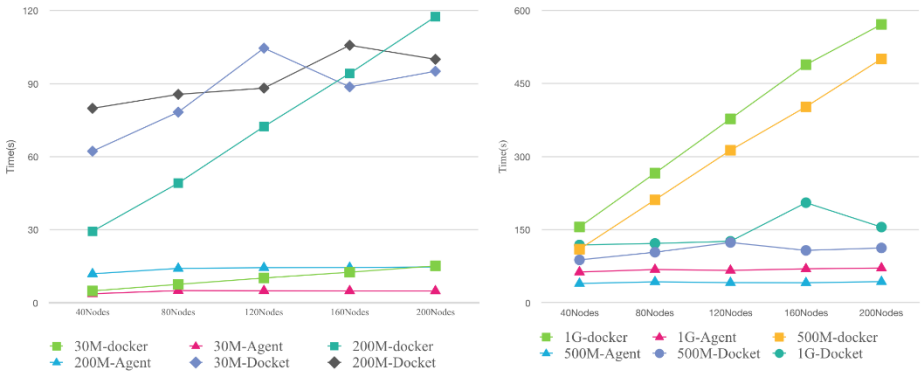
- **P2P Registry**：镜像仓库，种子生成，种子下载，文件初始提供者
- **Agent**：Docker透明代理，下载任务的主要功能组件
- **BT Tracker**：P2P下载资源查询定位组件



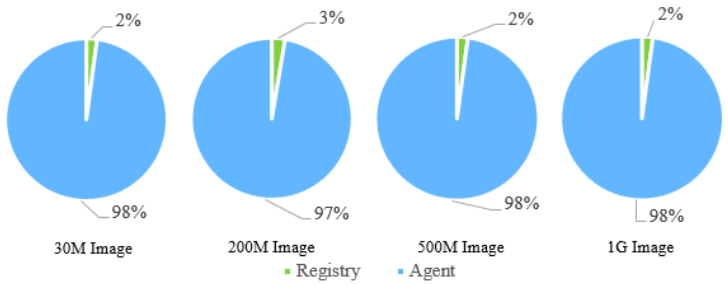
社区版 k8s在腾讯的演进— P2P Docker registry

- 每层分别做种，下载速度更快
- 优化流量调度算法，节省registry流量
- 代理方式，对Docker Daemon零入侵

Docker、Docket、Gaiastack P2P Agent下载镜像对比



Registry与P2P Agent流量占比对比



内容

1. GaiaStack平台简介

2.社区版 k8s在腾讯的演进

3. GaiaStack应用案例

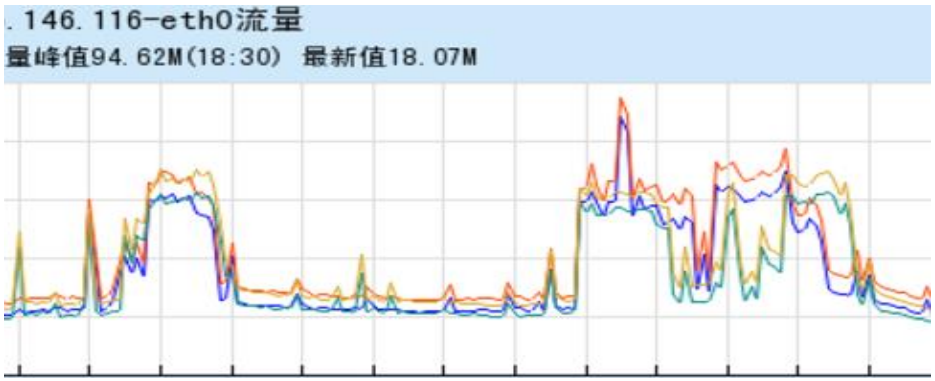
GaiaStack应用案例—腾讯信鸽

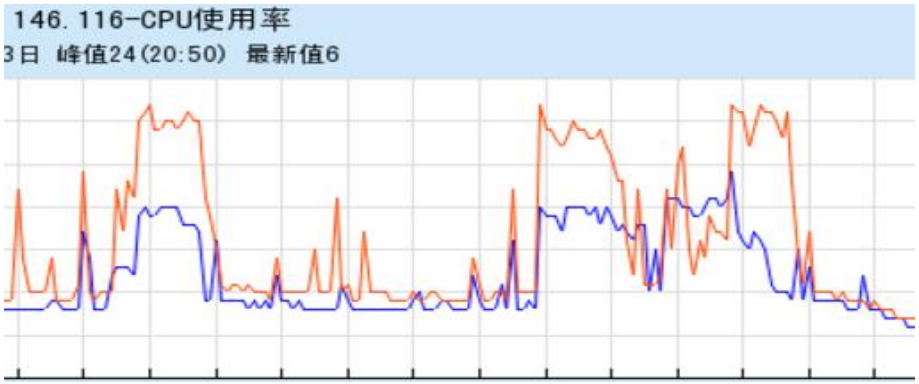
- 腾讯信鸽是国内领先的消息推送服务厂商。实现了亿级推送**10分钟内**完成，有效支持用户业务需求，累计服务数十万APP开发者。其中王者荣耀、穿越火线、快手、天天斗地主、分期乐、京东金融、KEEP、微店等众多知名APP都是信鸽平台的忠实用户。

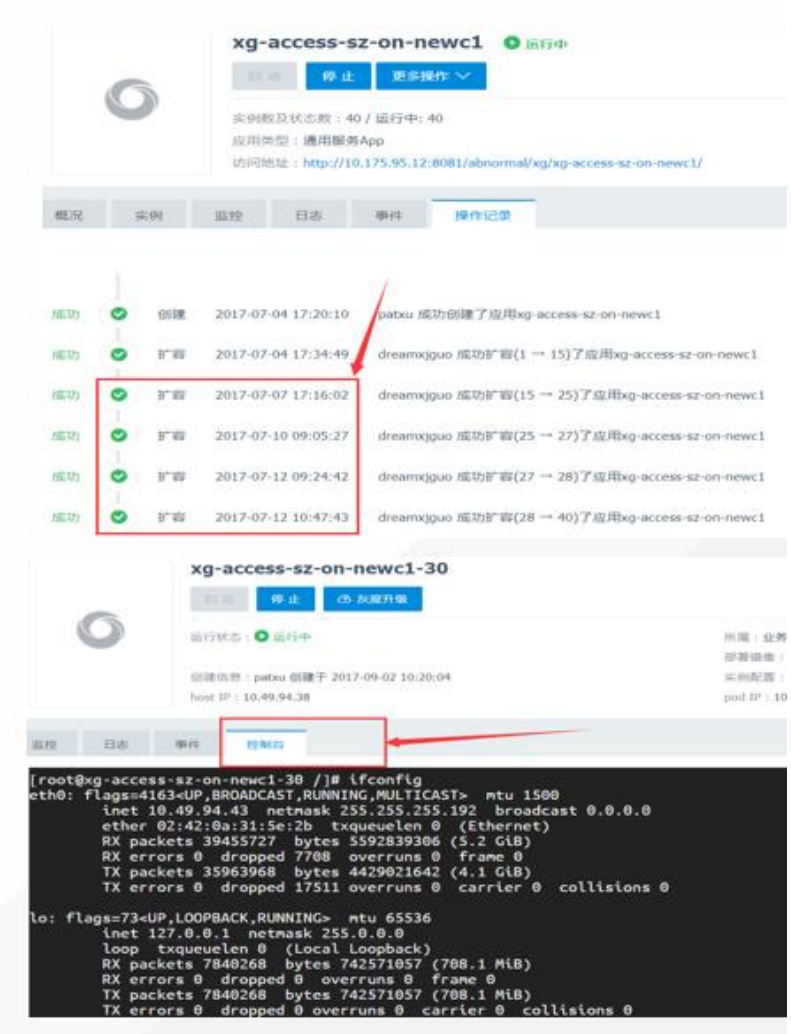
- GaiaStack解决的业务痛点
  - ✓ 多集群管理
  - ✓ 所有组件docker化
  - ✓ 混合部署、资源隔离
  - ✓ AutoScaling，自动加入TGW、L5
  - ✓ Floating IP
  - ✓ 网络层启用 SR-IOV 特性

GaiaStack应用案例—腾讯信鸽

docker & SR-IOV，高峰 CPU 降低约 38.3%







Thank you !