十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店

腾讯GaiaStack容器产品私有云场景实践

腾讯 陈纯

自我介绍

陈纯，腾讯TEG数据平台部高级工程师

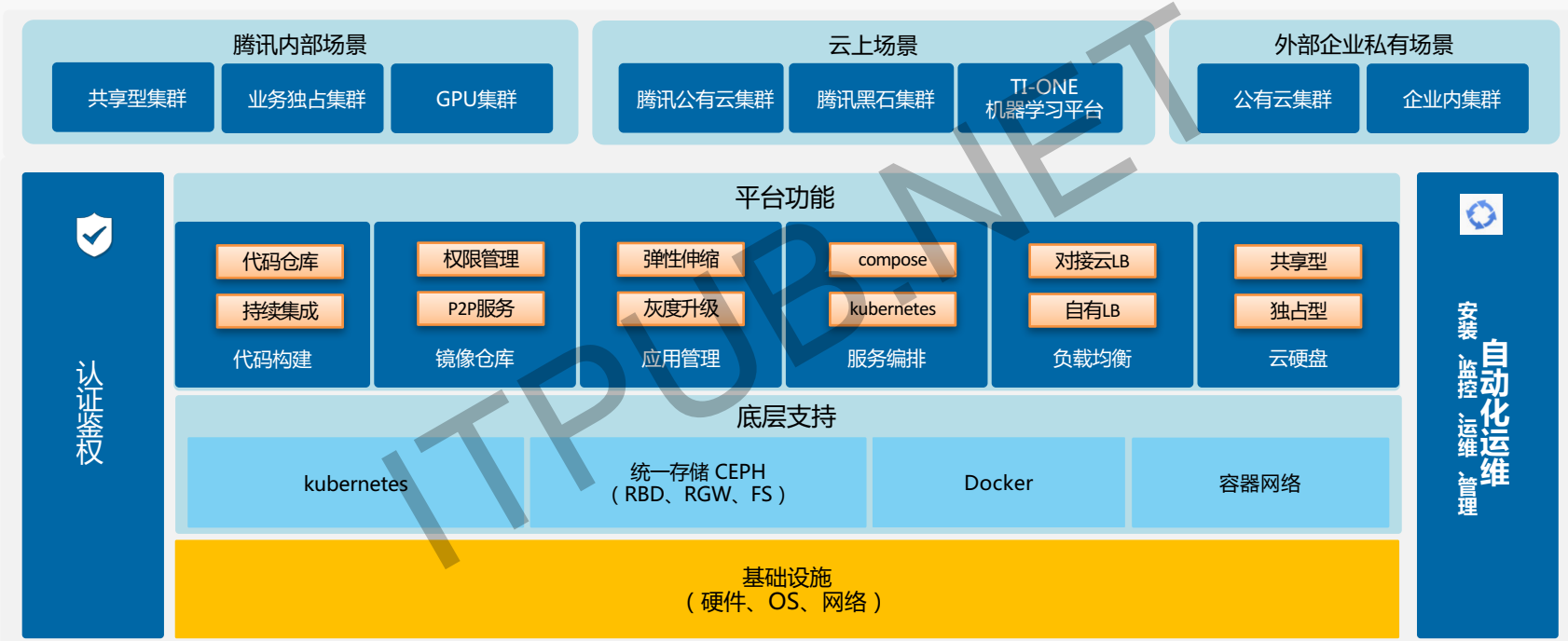
- Docker Libnetwork项目maintainer
- 《循序渐进学docker》作者之一
- 对容器生态的Kubernetes, Docker, runc, Libnetwork, Flannel等开源项目都有大量的源代码贡献

GaiaStack介绍

GaiaStack是腾讯基于Kubernetes打造的企业级容器云平台。

GaiaStack作为数据中心操作系统，可以调度CPU或GPU等海量计算资源，运行所有计算框架，监控任务执行结果，让数据中心的所有资源被合理使用，协同完成各类应用场景的计算。

GaiaStack功能及应用场景

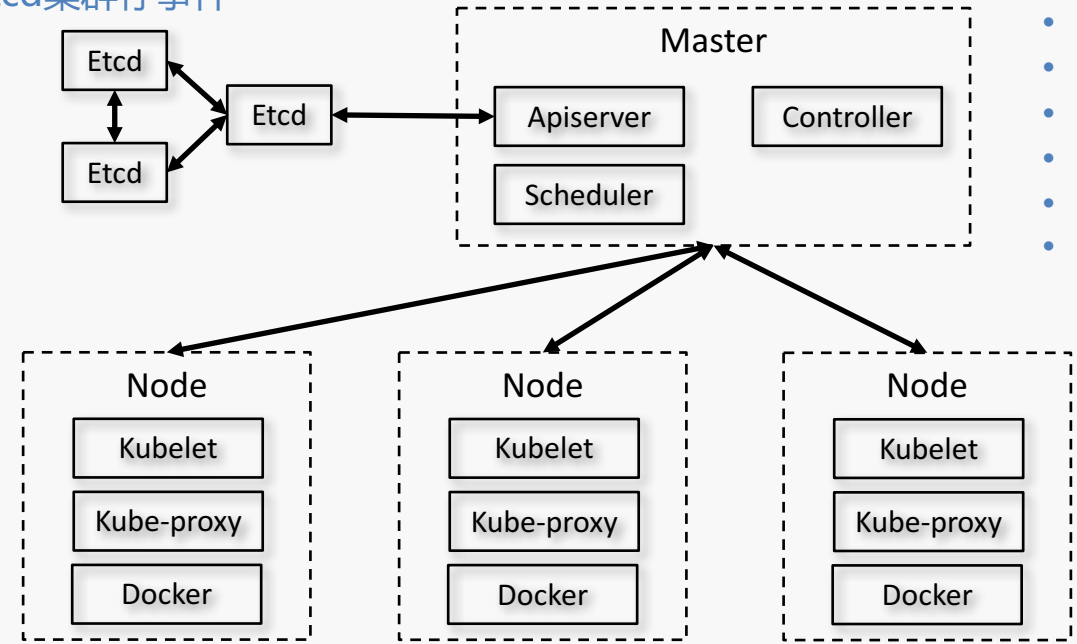


社区版Kubernetes私有云场景挑战

- 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

Kubernetes扩展接口

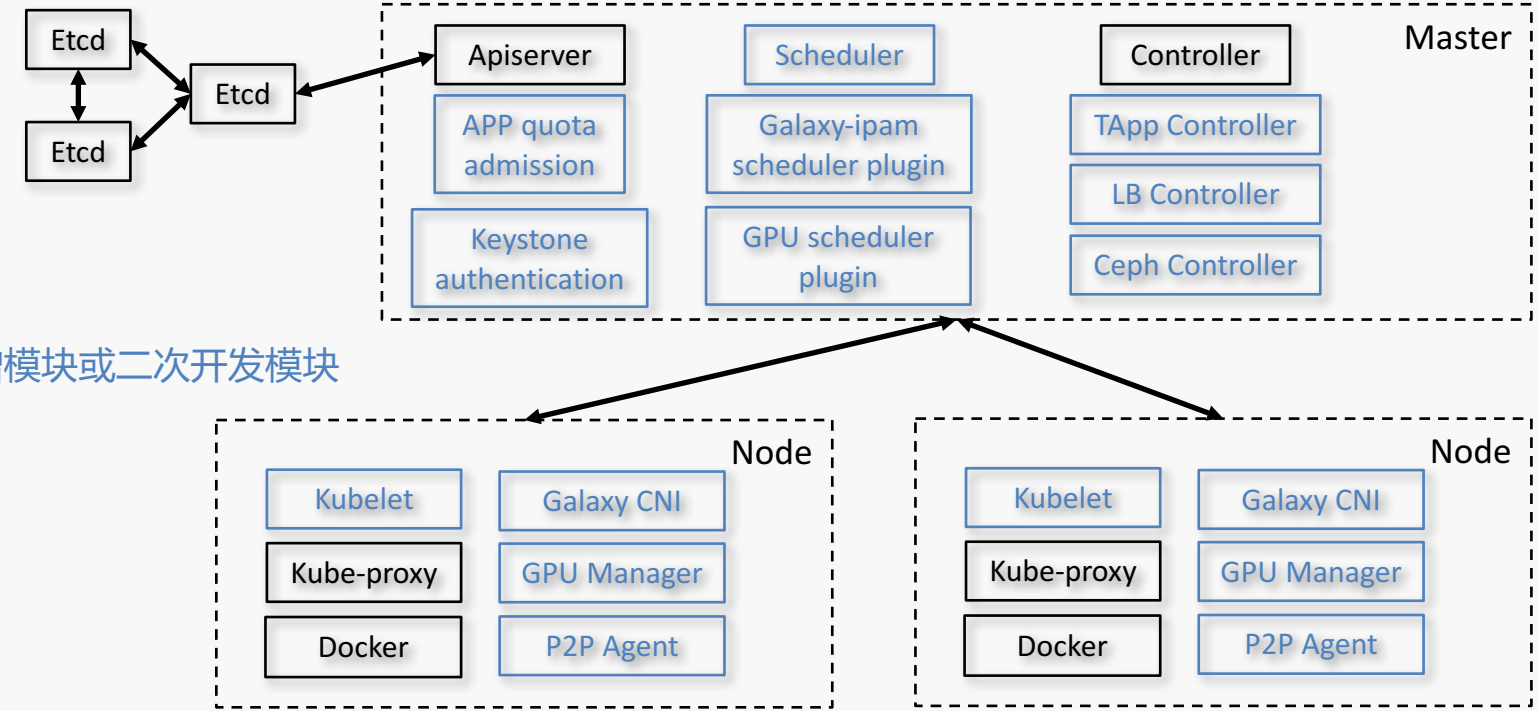
单独etcd集群存事件



- 自定义API类型CRD
- 认证、授权、准入扩展
- 多个Apiserver存不同对象
- 调度器插件
- 多调度器
- 自定义控制器

- 存储插件接口CSI
- 设备插件DeviceManager
- 网络插件CNI
- 容器运行时接口CRI
- Kube-proxy替代项目

GaiaStack Kubernetes



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

优势：

- **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
- 不只是修改镜像版本，甚至可以多加一个容器

```
type TAppSpec struct {
    Template corev1.PodTemplateSpec //默认Pod模板
    TemplatePool map[string]corev1.PodTemplateSpec
    Statuses map[string]InstanceStatus
}
```

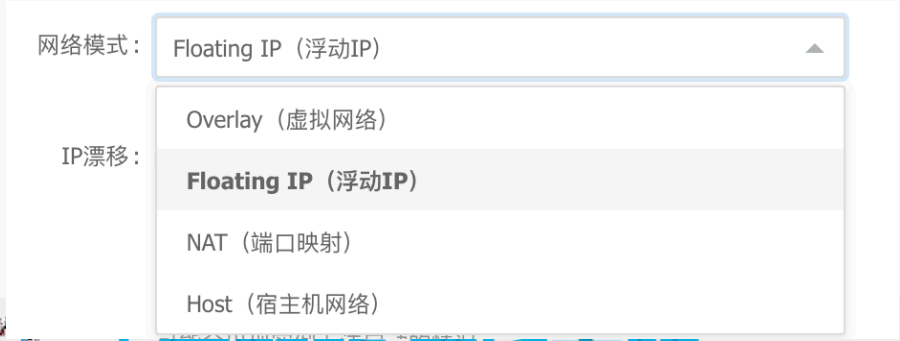
Galaxy-CNI网络插件

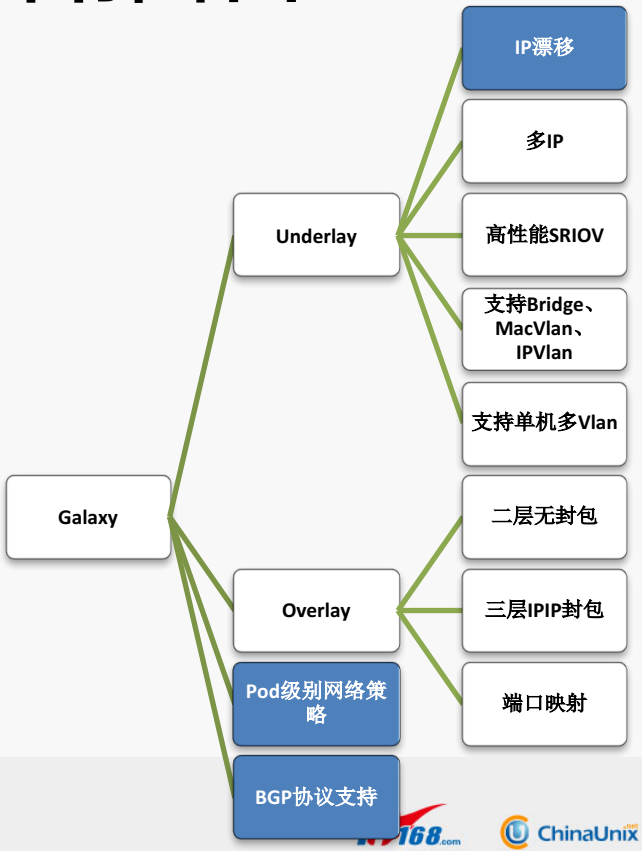
因为各种不同场景的需要，自研网络项目Galaxy

- 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

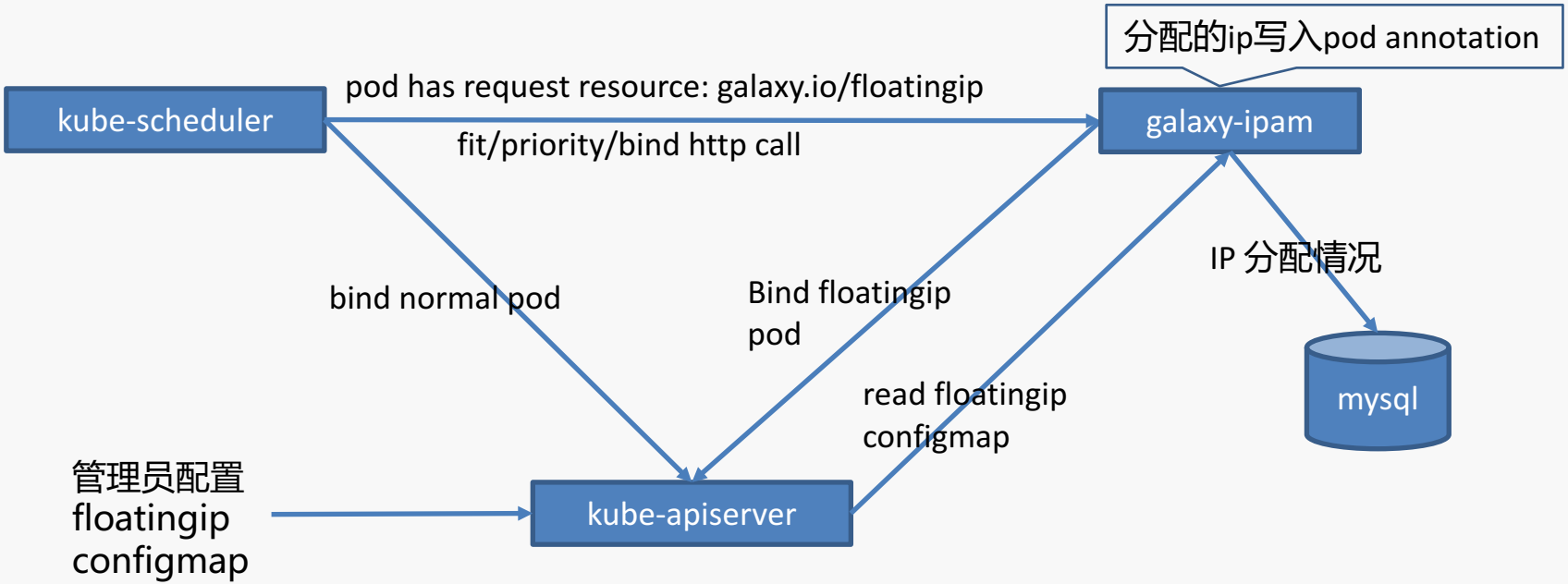
选择应用适合的网络方案

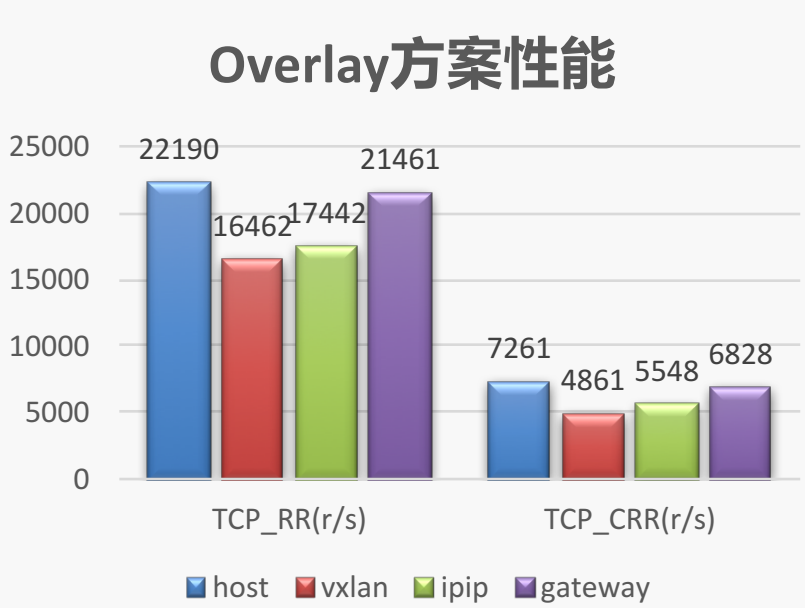
- 不同的应用可以选择不同的网络模式
- **同一主机的不同容器可以选择不同的网络模式**

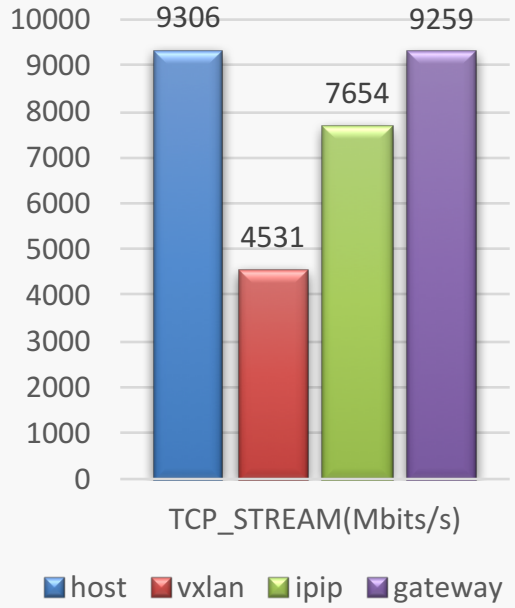




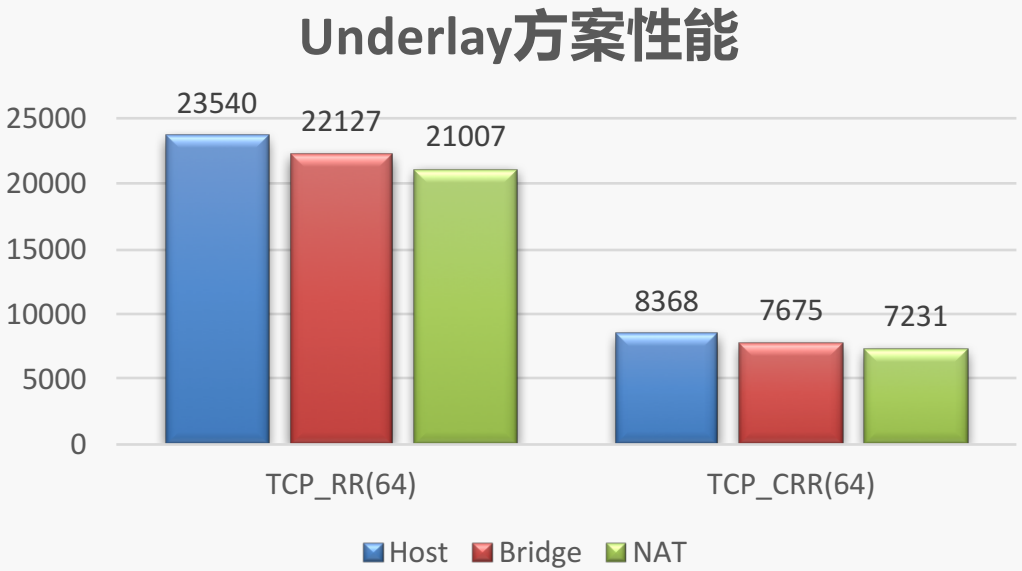
Pod Underlay IP漂移





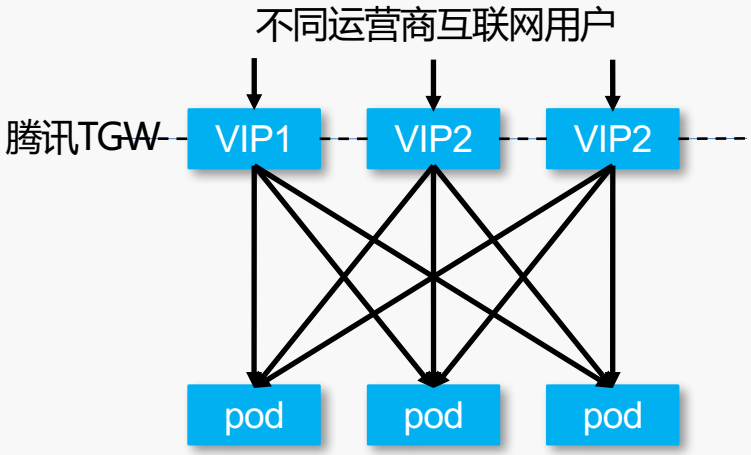


GaiaStack提供Overlay方案是IPIP + Host Gateway 混合方案，性能仅次于calico3层方案
短链接 Vxlan比HOST**差33%**，IPIP比HOST**差23%**，Gateway比HOST只差**6%**
方案被社区合并 https://github.com/coreos/flannel/pull/842



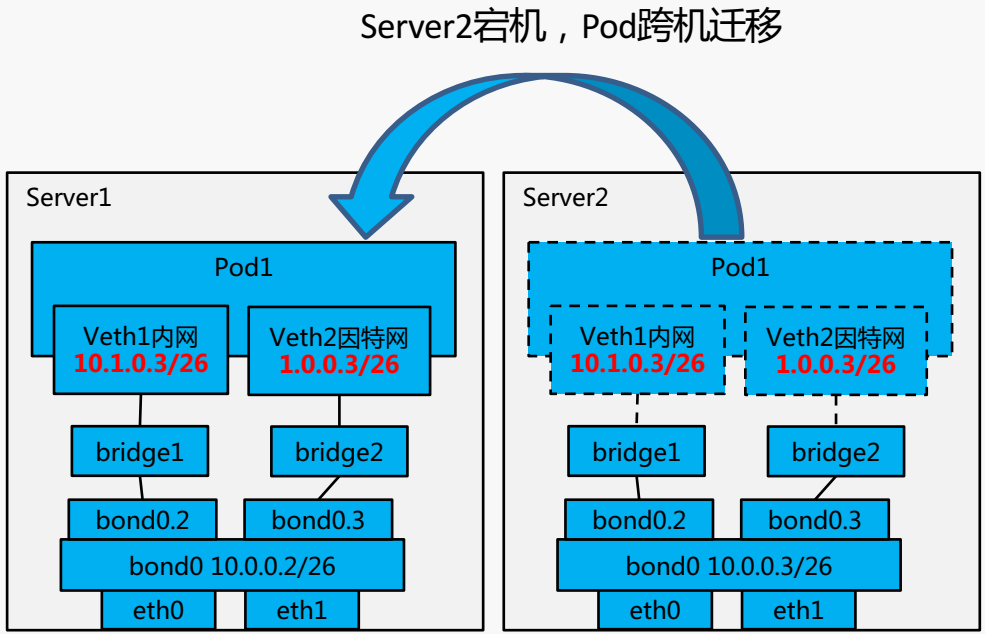
Bridge比Host仅**差6%**

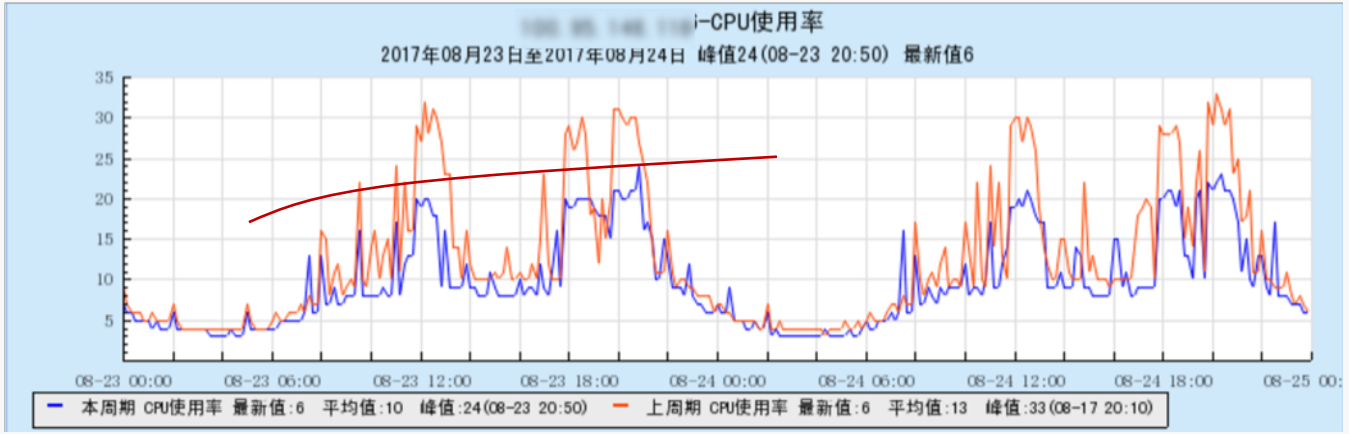
Underlay网络案例



Bridge/SRIOV Underlay网络

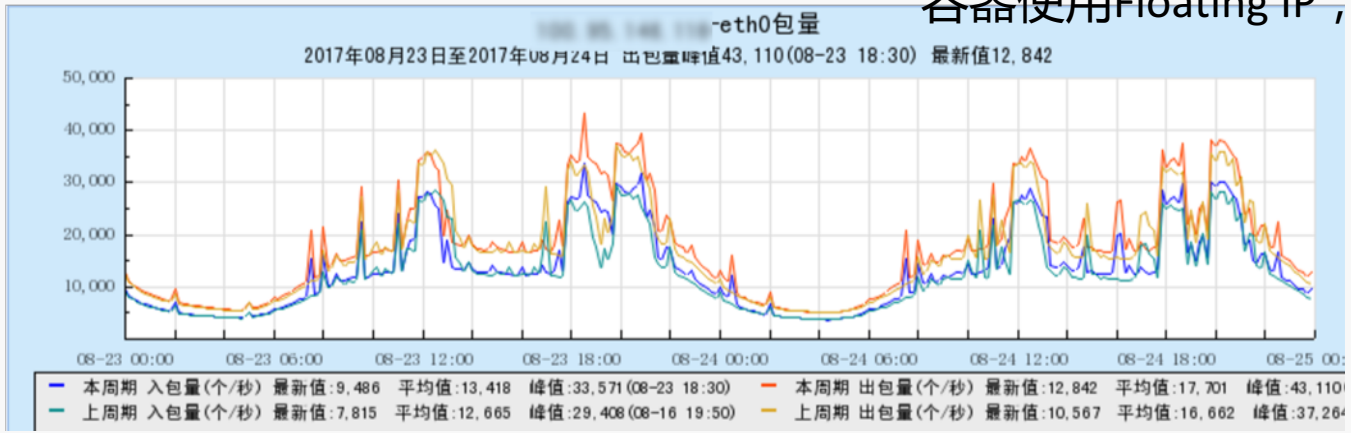
将容器的IP和端口直接绑定到TGW
将TGW的Tunnel网卡创建到容器中





CPU使用降低1/3

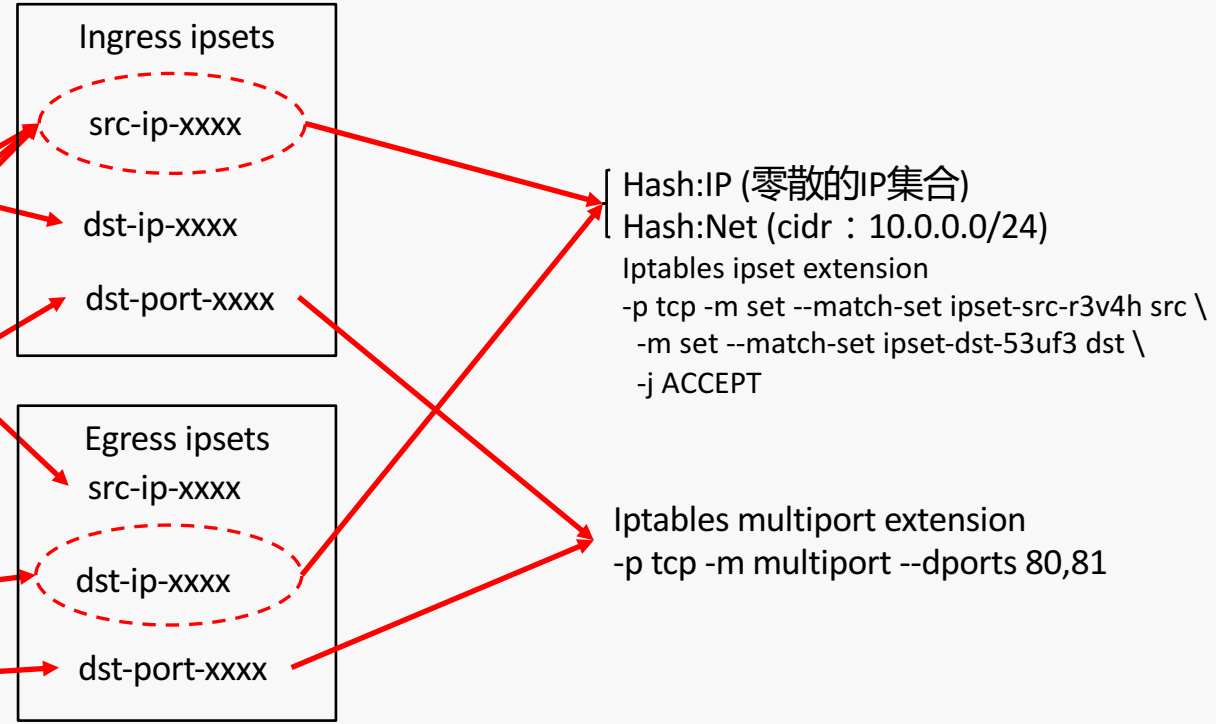
容器使用Floating IP，SRIOV网桥方案



包量+6%

实现kubernetes网络策略

```yaml
apiVersion: networking.k8s.io/v1
kind: NetworkPolicy
metadata:
  name: test-network-policy
  namespace: default
spec:
  podSelector:
    matchLabels:
      role: db
  policyTypes:
  - Ingress
  - Egress
  ingress:
  - from:
    - ipBlock:
        cidr: 172.17.0.0/16
        except:
        - 172.17.1.0/24
    - namespaceSelector:
        matchLabels:
          project: myproject
    - podSelector:
        matchLabels:
          role: frontend
    ports:
    - protocol: TCP
      port: 6379
  egress:
  - to:
    - ipBlock:
        cidr: 10.0.0.0/24
    ports:
    - protocol: TCP
      port: 5978
```



BGP协议的使用

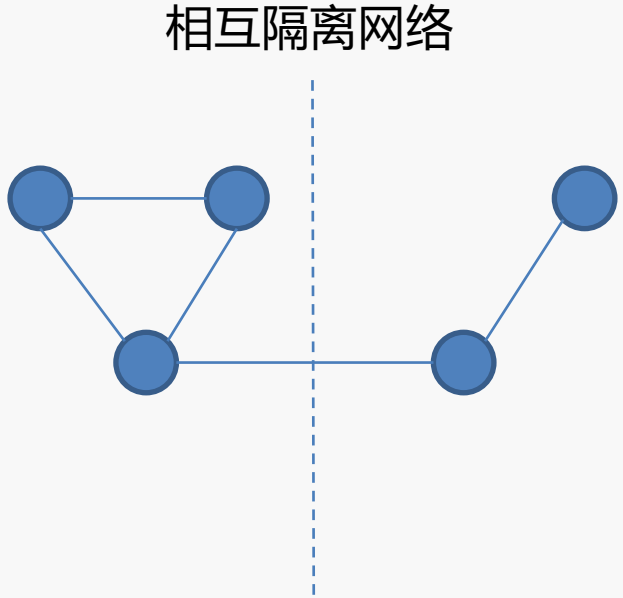
Calico/contiv提供了BGP L3 network的实现，给容器提供一个完全无封包的网络

- 要求交换机开启BGP功能，实质是一种基于BGP协议的数据中心网络架构。

BGP对于容器网络的价值：

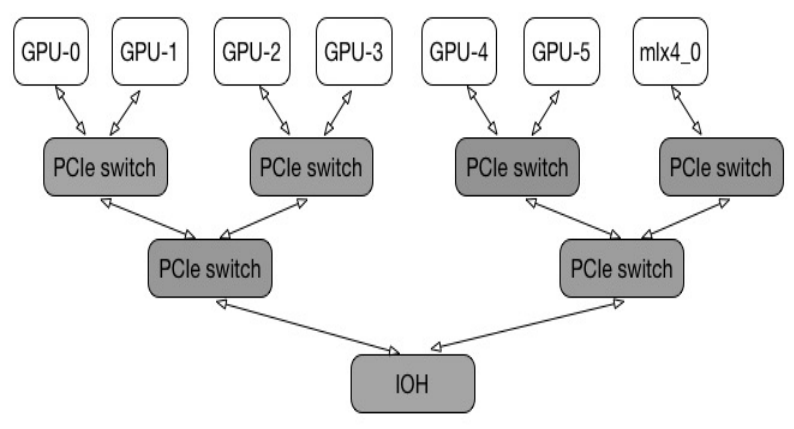
- 无封包的网络方案
- 解决不完全连通网络的容器跨主机访问的问题。相比Weave用gossip协议实现P2P网络，标准路由协议的优势：AS Path防环路；Route Reflector/Route Server解决P2P的性能问题

Galaxy借助GoBGP实现BGP协议的支持



GPUManager设备插件

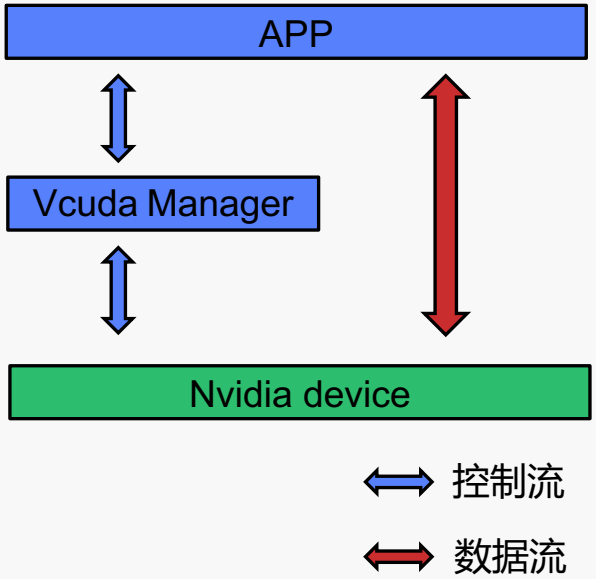
- 自研Kube-scheduler调度器插件 + Device Plugin GPU Manager
- 异构集群的精细化管理
  - 管理不同型号的GPU（如：M40、P40等）组成的集群限定不同业务可以使用的卡种类和数量
  - 保证业务能够使到预期的GPU资源（GPU卡数，GPU型号）
- 调度分配优化
  - Drain node调度解决资源碎片问题
  - 多卡需求按GPU架构优先分配距离较近的卡

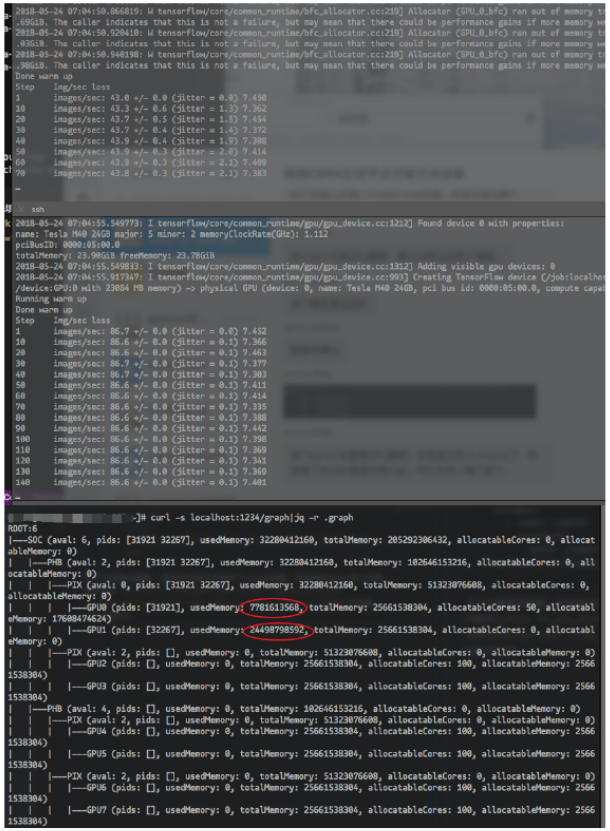


GPU虚拟化

业内首创容器场景的GPU虚拟化技术

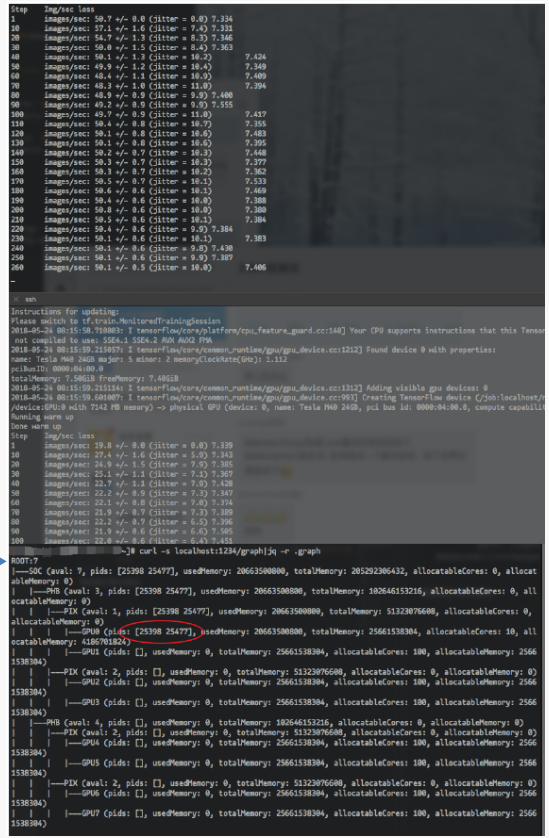
- 多个程序共享同一张GPU卡，提高资源利用率
- 支持GPU卡，GPU内存的虚拟
  tencent.com/vcuda-core=0-100、
  tencent.com/vcuda-memory=0-100
- GPU卡和内存资源申请通过Manager进行限速，用户程序的数据流不经过Manager
- 对用户程序零入侵
- 基于kubernetes device plugin实现，对kubernetes、 kernel无入侵





vcuda-1容器器 申请了了50%的利用率，7680MB，运行行行在0号GPU，
vcuda-2容器器独占卡，
运行在1号GPU；vcuda-1的训练速率是平均43.8/s，vcuda-2的训练速度是平均86.6/s

vcuda-1容器器 申请了了30%的利利用用率，7680MB，运行在0号GPU，
vcuda-2容器器申请了60%利用率，12800MB，运行在0号GPU；vcuda-1的的训练速率是平均25.22/s， vcuda-2的训练速度是平均54.7/s

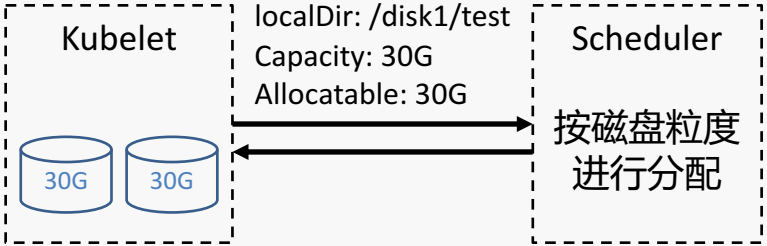


磁盘空间管理

GaiaStack修改了调度器和Kubelet以支持本地磁盘管理，相比Kubernetes emptyDir优势：

- 支持多块盘
- 支持弹性管理，允许临时超出
  - 单个Pod超出limit大小后若机器磁盘仍有余量可以暂时使用

目前耦合在Kubernetes代码中，后续计划用scheduler plugin和CSI接口进行优化

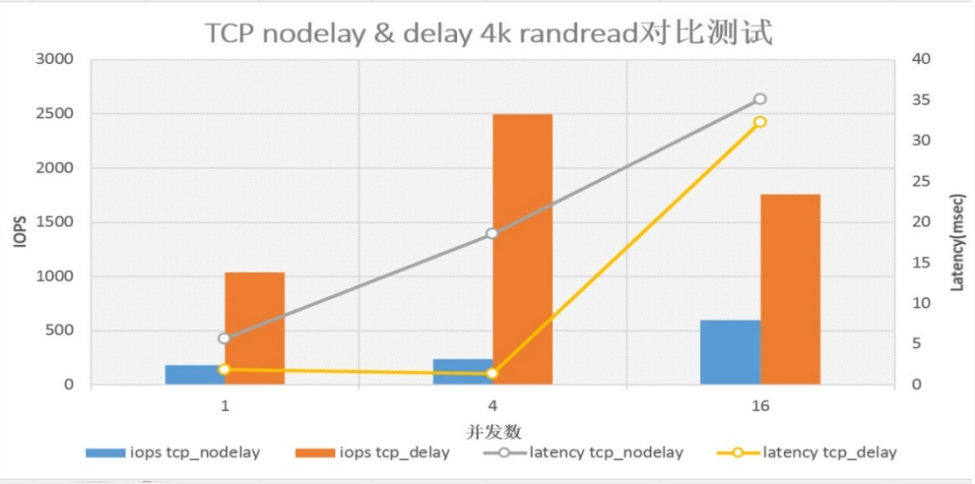


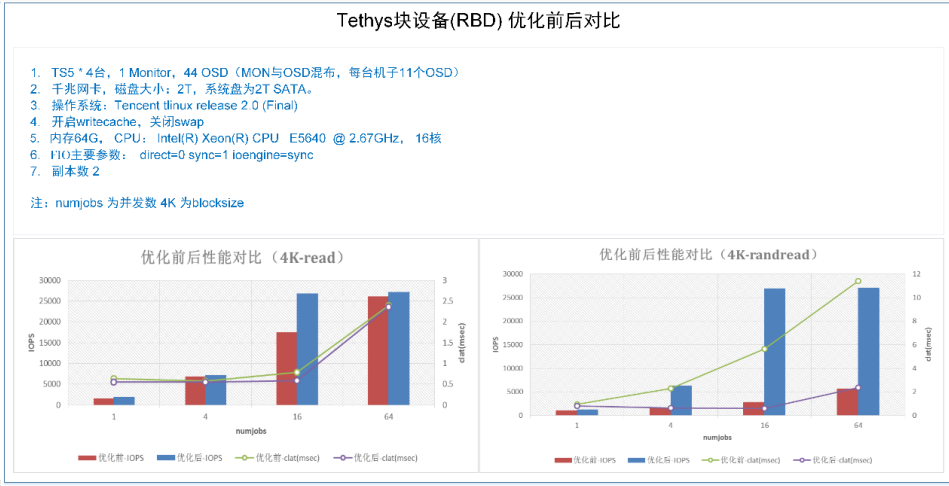
云盘管理

- 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

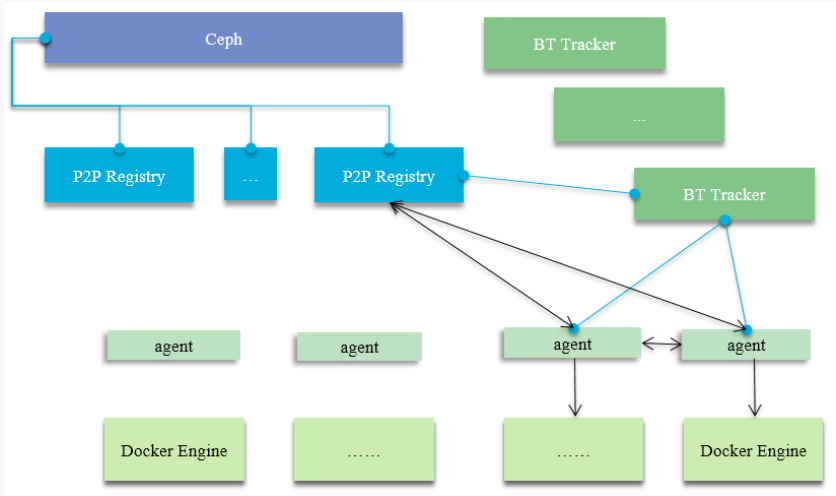
一名Ceph官方组织成员

- 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





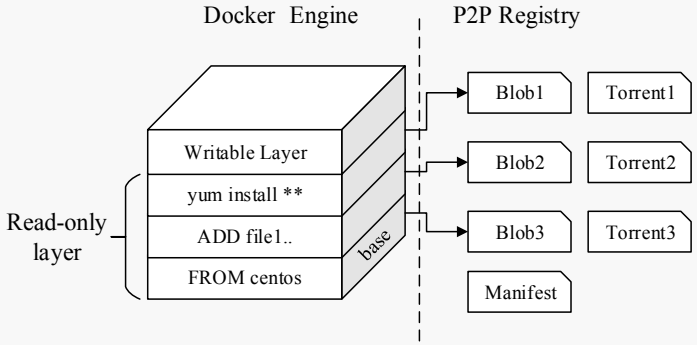
P2P Docker registry



- **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

主要设计思想：

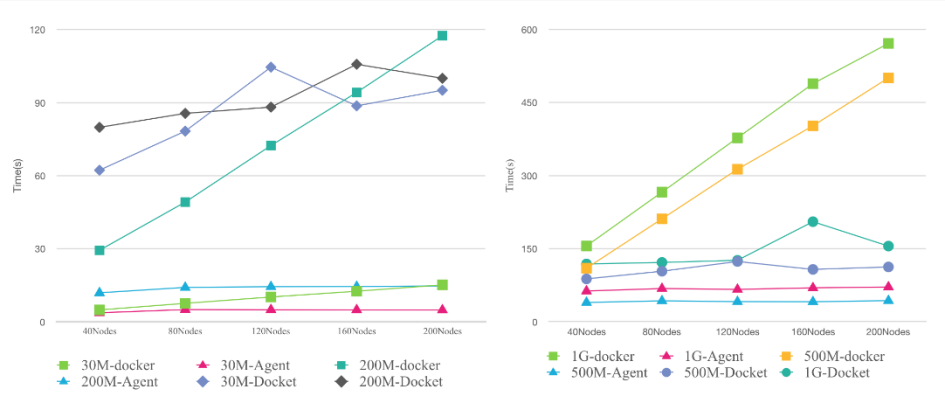
- 在镜像下载过程中，引入BT协议
- 在Blob上传时，对Blob生成种子
- 每层分别做种
- 在下载镜像的Blob时，先下载种子，再通过种子文件下载数据



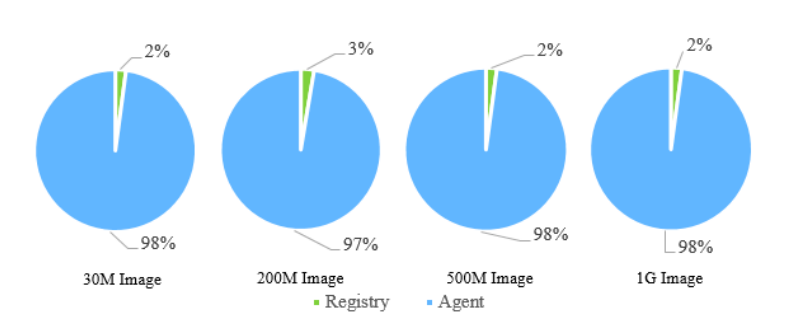
P2P Docker registry

- 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



热升级问题

1.4升级1.9版本

- Pod Hash发生变化
- Container名称发生变化，点分隔改为了下划线分隔
- 容器标签发生变化
  - pause容器的标签io.kubernetes.container.name=POD改为io.kubernetes.docker.type=podsandbox
  - io.kubernetes.container.restartCount改为annotation.io.kubernetes.container.restartCoun
- Cgroup目录结构发生变化，新增Pod层级

THANKS