云原生社区
Cloud Native Community

SHANGHAI

云原生社区Meetup第一期

主办方 云原生社区 Cloud Native Community　承办方 UCLOUD优刻得 云原生社区上海站 Cloud Native Community Shanghai　赞助商 tetrate 万科海外 Broadview

2020-11-28　上海站

高鹏 UCLOUD优刻得

UCloud后台研发工程师

UCloud 后台研发工程师，负责内部云原生平台的建设。在 Kubernetes 的落地实践、高可用架构的设计等方面拥有丰富的经验。

云原生社区Meetup
第一期·上海站 × UCAN 技术沙龙

云原生社区 Cloud Native Community

UCLOUD优刻得

Kubernetes 在 UCloud 内部的应用

演讲人：高鹏

UCLOUD优刻得

目录

KUN | 鲲

北冥有鱼，其名为鲲。鲲之大，不知其几千里也。

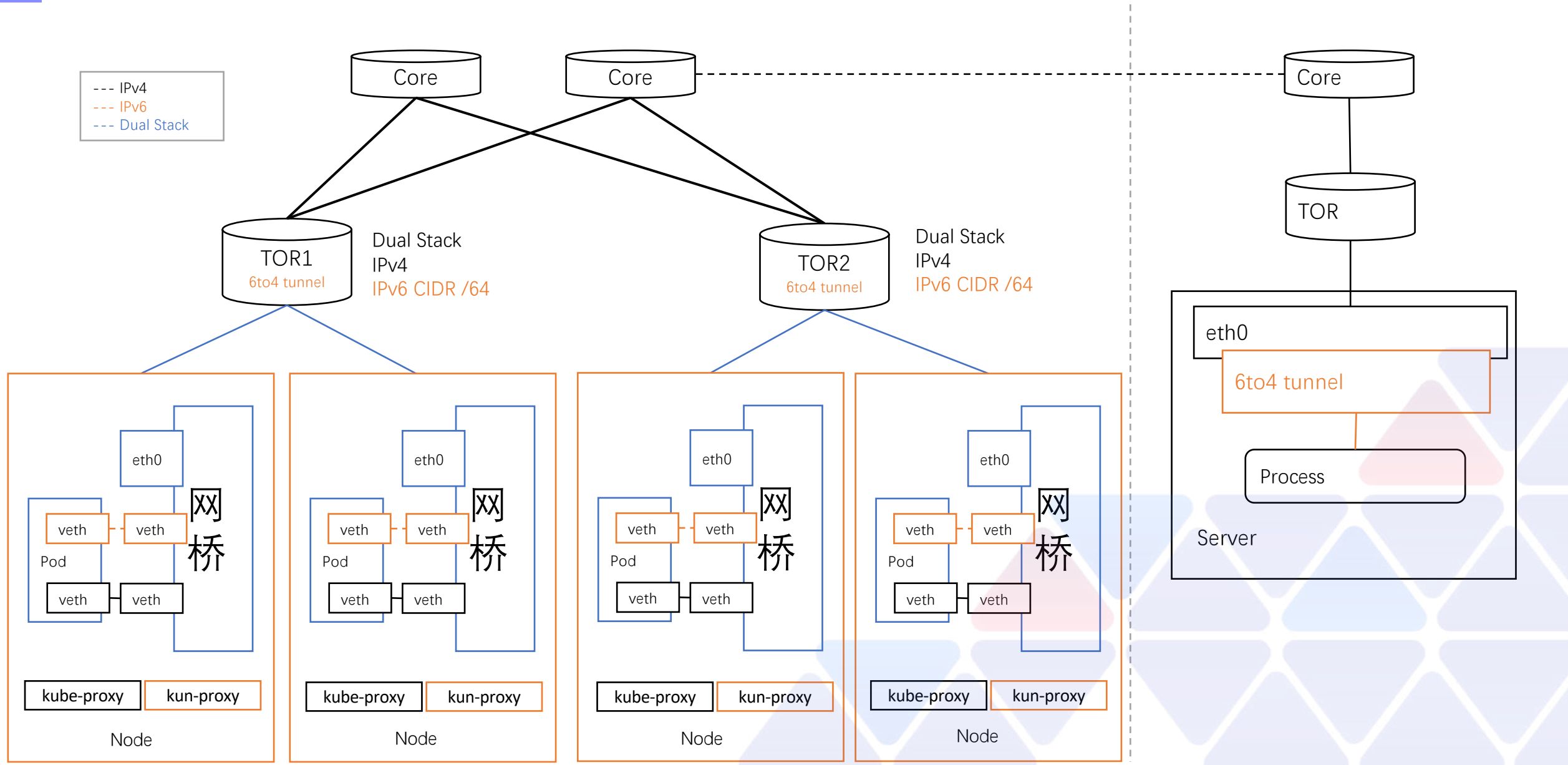
鲲项目基于 Kubernetes，面向 UCloud 内部，通过一系列的平台和工具，帮助您更快地研发产品、提供更好的服务。

面向内部的 Kubernetes

网络背景

- IPv4 地址不够用
- Kubernetes 集群内部的 Pod 和 Service 需要和集群外部互通
- 需要通 VPC

IPv6/IPv4 双栈网络方案 V1



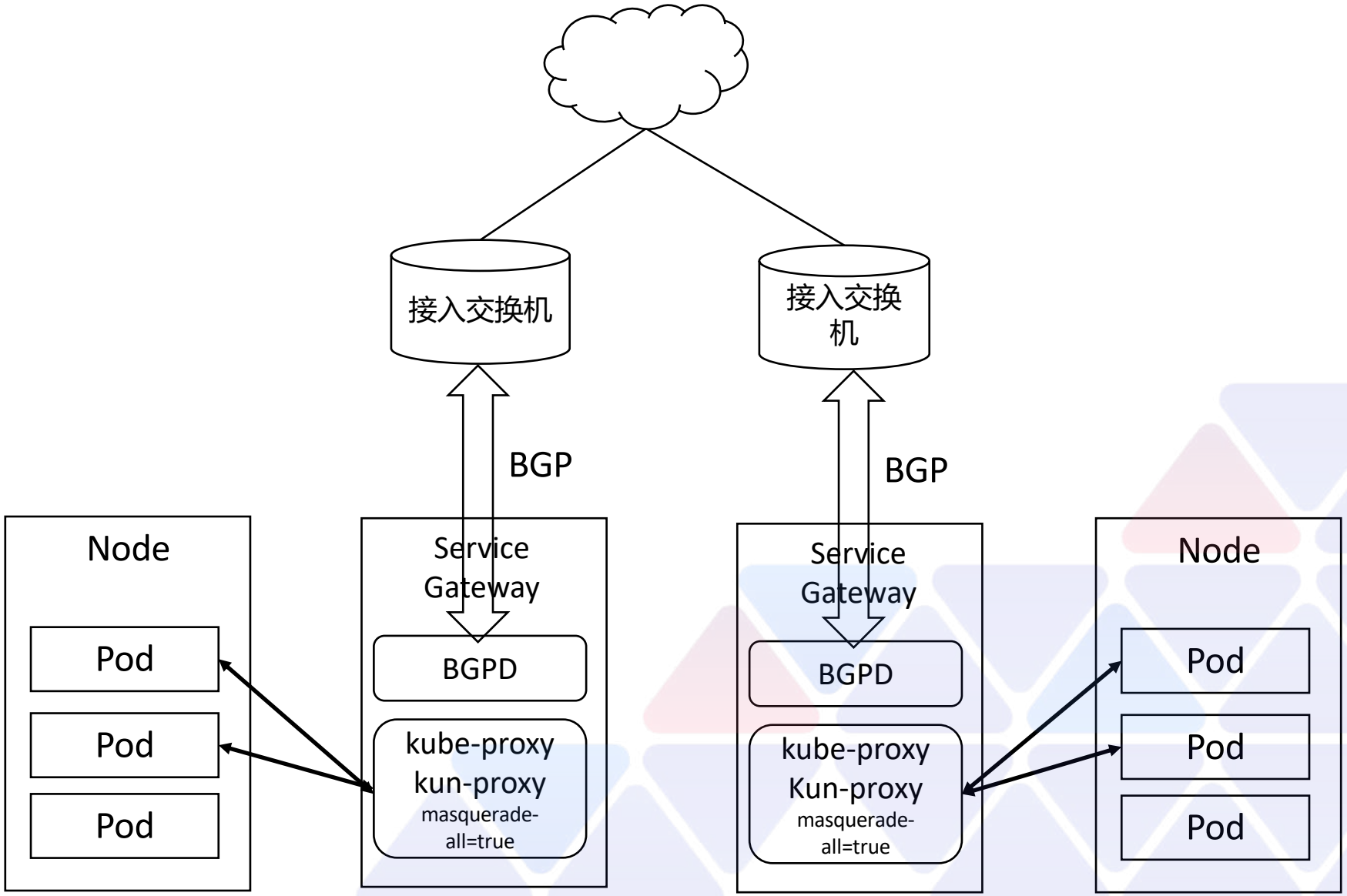
IPv6/IPv4 双栈网络方案 V1

方案

- 每个机房的每台交换机下 2+ 台网关
- 接入交换机宣告 IPv4 VIP
- 网关和交换机使用 BGP 实现 ECMP，宣告 VIP 对于的 6to4 IPv6 地址
- Kube-Proxy/kun-proxy 实现 SNAT

特性

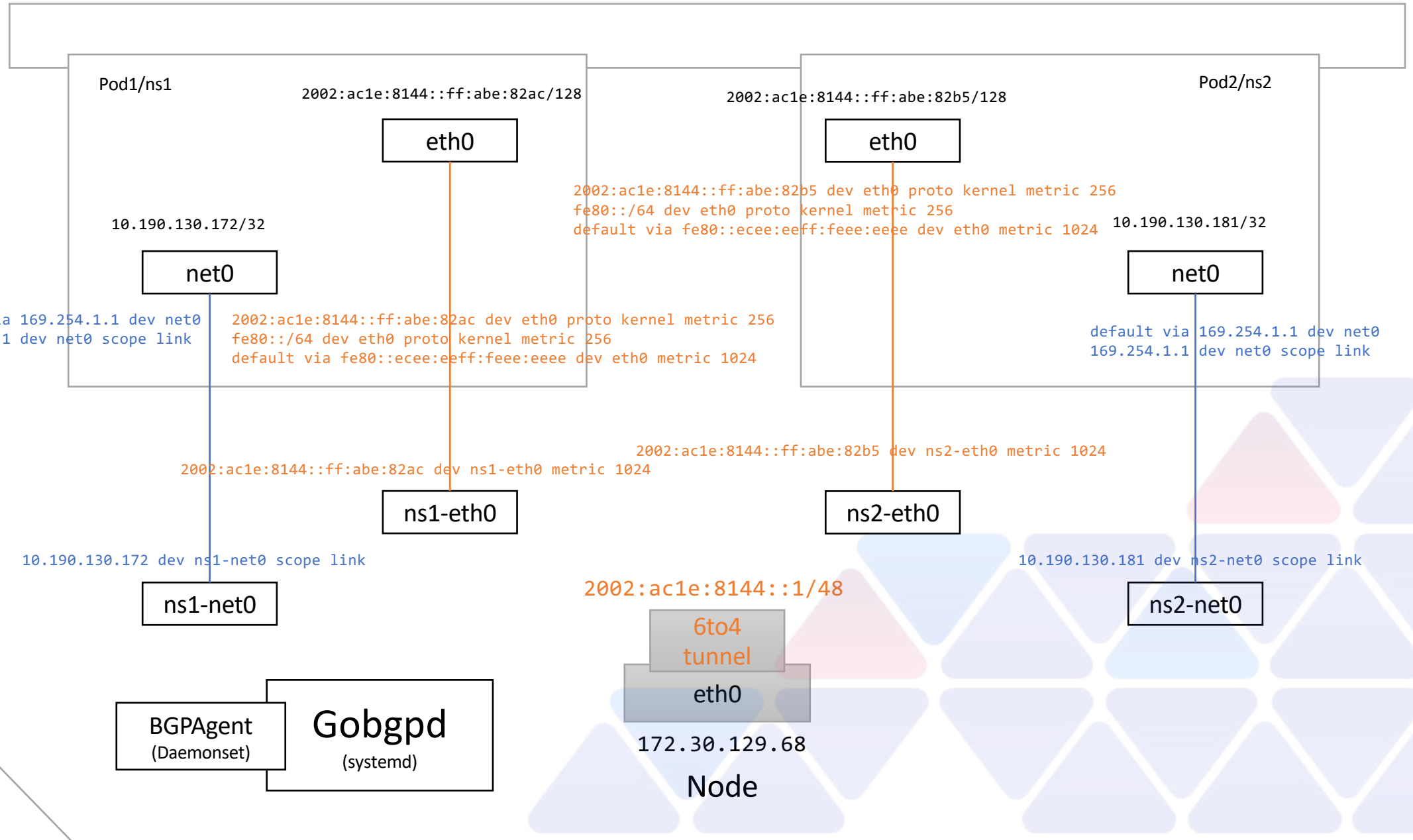
- ClusterIP 本身集群内网都可访问
- 跨机房高可用



问题

TOR 需要特殊配置支持 IPv6 隧道，扩容、运维成本较大

IPv6/IPv4 双栈网络方案 V2



基于 Calico BGP 6to4 隧道下沉到服务器

CRD

- BGPPeer
- BGPRouteAffinity

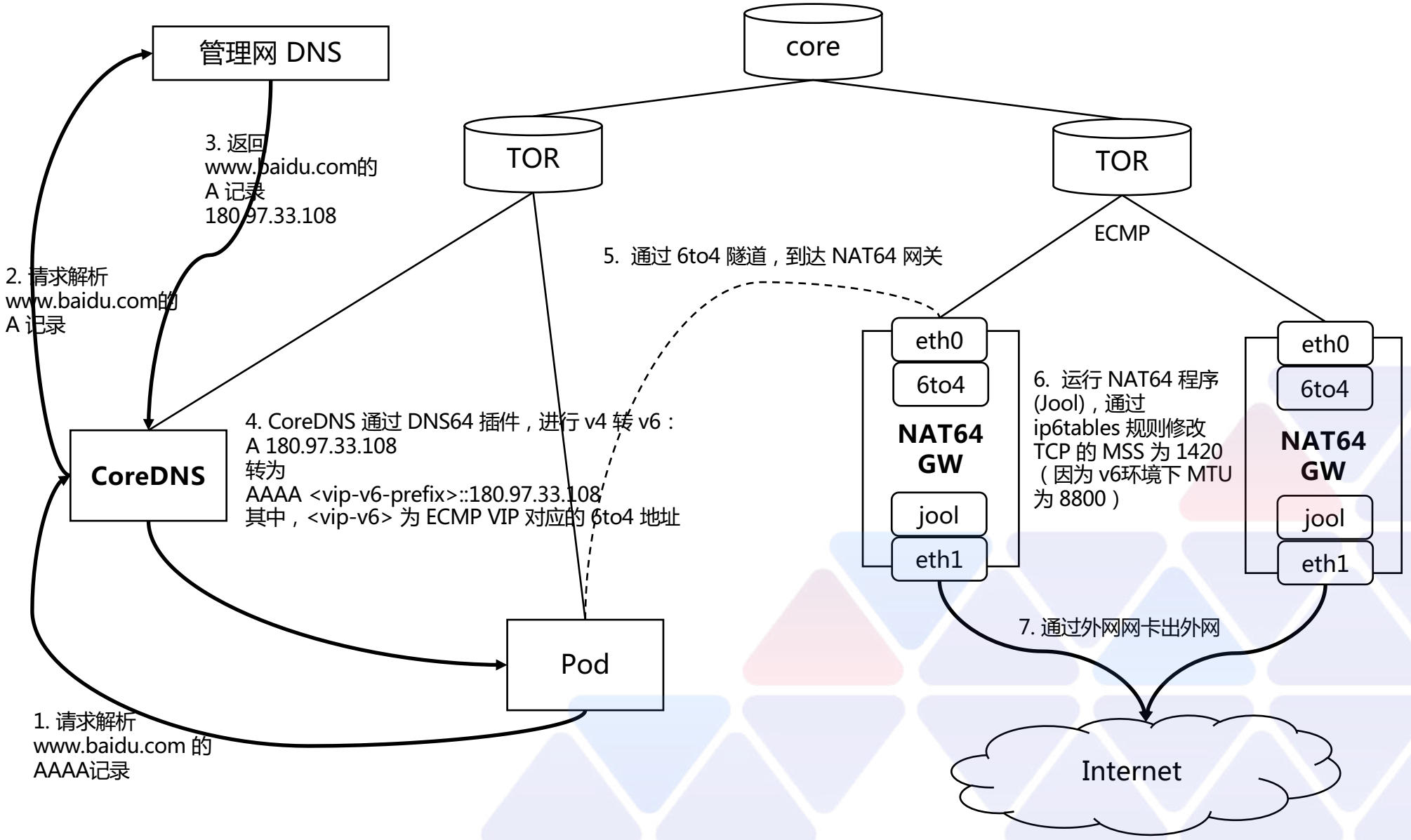
访问外网

方案

- NAT64(Jool)
- DNS64(CoreDNS Plugin)
- ECMP（BGP）

特性

- 高可用
- 科学上网



网络的一些补充

- Calico 的改造
  - IPAM 可能会出现 IP 短时间内重复利用的问题，我们做了修改，增加了 IP “冷却时间”
  - 每个 Pod 可创建 2 对 Veth Pair，分别用于 IPv6 和 IPv4，并且做了映射，IPv6 地址的最后 32 位即为 IPv4 地址
  - 使用 gobgpd 进行 BGP 通信
  - 增加了 Pod Annotation 开关控制是否需要 IPv4
- Kubernetes 双栈支持不太好
  - Kube-proxy 镜像中 conntrack 版本太老，清除 IPv6 NAT Entry 会有问题，我们升级到了conntrack v1.4.5
  - Kube-proxy 在 IPVS 模式时，其依赖的 libnetwork 库一些地方写死了 ipFamily 为 IPv4，k8s v1.18.9 后修复。我们使用 iptables 模式。
  - 当 svc 和 endpoints 数量较多时，iptables 更新会有较大延迟（10s），需要业务加入 preStop 步骤，进行等待。
  - K8S 从 v1.16 才开始提供双栈支持（alpha），kubernetes v.16 和 v.17 版本对 iptables 模式的双栈 service 支持不友好，不可以同时支持 IPv6 和 IPv4 的 service，只能支持其中一种，ipvs 模式的双栈 service 则没有此限制；kubernetes v1.18及其以后的 iptables 模式的双栈 service 可以同时支持 IPv6 和 IPv4。
  - 我们开发了 kun-proxy（基于 IPVS）、kunservice（CRD）以及 kunservice operator，和 kube-proxy 以及 service一起工作，分别支持 IPv4 和 IPv6。同时实现 AZ 亲和性。
  - 开发了 CoreDNS 插件提供 6to4、NAT64、科学上网的功能

存储

- 块
  - Ceph RBD on Rook，混用 K8S Node
    - Rook v1.0 之前在线升级有问题，我们进行了改造支持优雅升级
    - cgroups 内存限制下 osd flapping 问题
  - 云硬盘 UDisk
- 文件
  - UFS(NFS)
    - 为提升高可用能力，我们开发了基于 ipvs 的 NFS 高可用客户端
  - US3(s3fs mount)
- 对象
  - US3

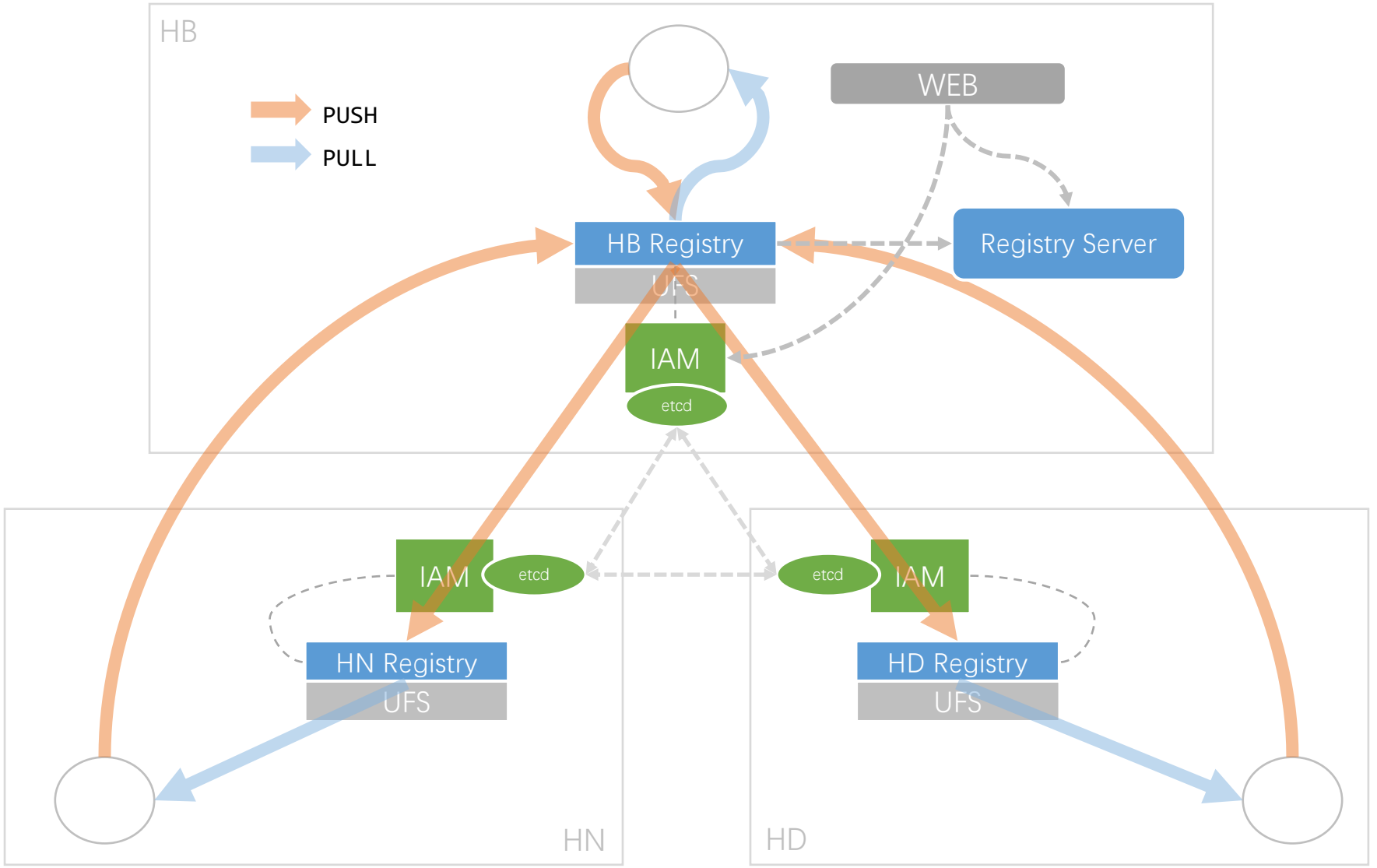
镜像仓库

方案

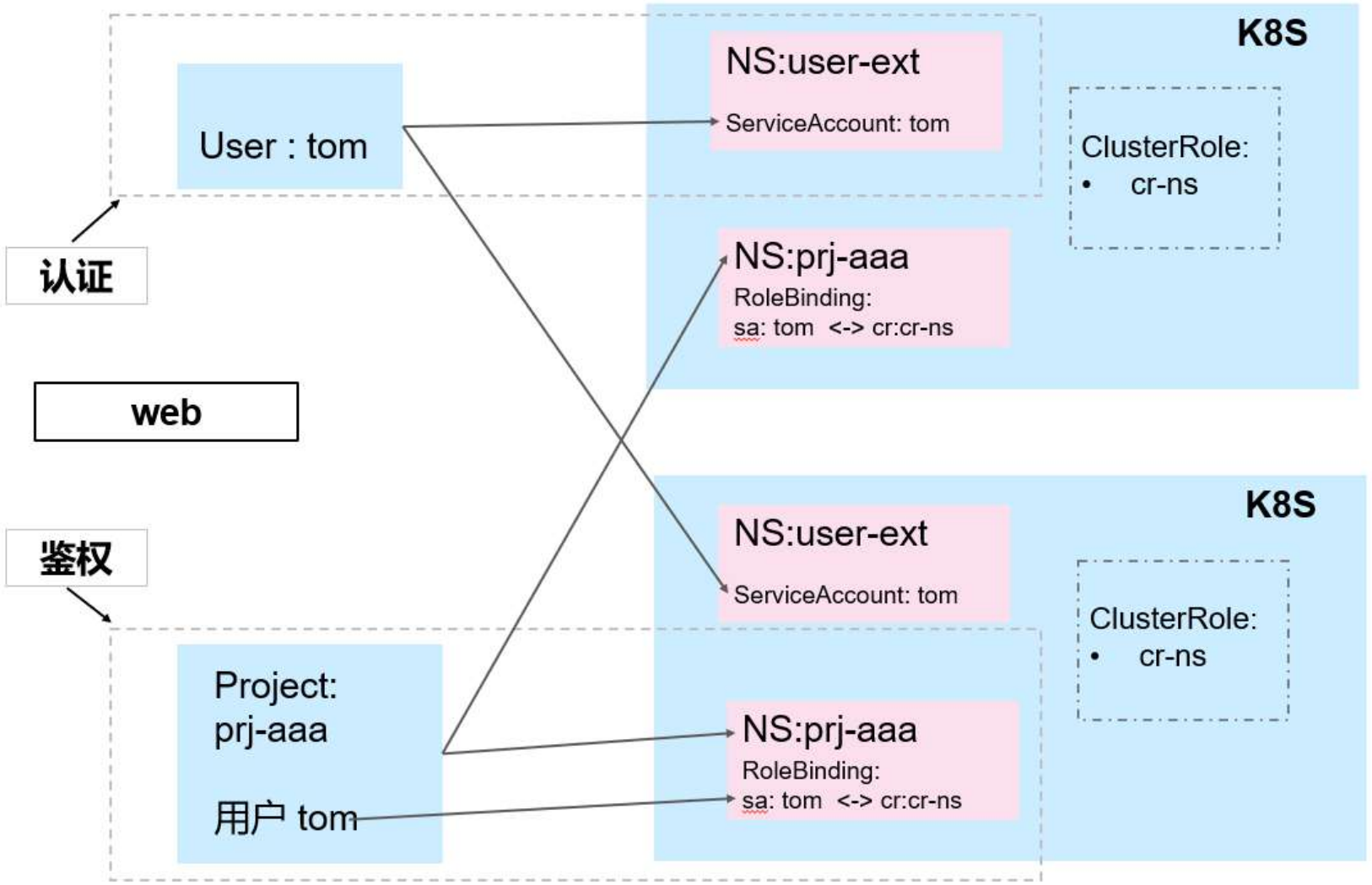
- 多地域部署，一主多从
- 权限系统基于多地部署的 etcd
- 存储使用 UFS，在客户端用 Keepalived 实现 UFS 高可用
- ECMP + Nginx 实现 L3/L4 的 HA 和 LB
- Storage Driver，Push 时主从同时写，Pull 时就近选择服务，读本地数据
- Registry Server 支持镜像列表的查询和操作

特性

- Region 级别容灾，发生Region不可用时，服务降级，镜像能正常拉取



多租户


K8S
NS:user-ext
ServiceAccount: tom
ClusterRole:
• cr-ns
User : tom
NS:prj-aaa
RoleBinding:
sa: tom <-> cr:cr-ns
认证
web
鉴权
K8S
NS:user-ext
ServiceAccount: tom
ClusterRole:
• cr-ns
Project:
prj-aaa
用户 tom
NS:prj-aaa
RoleBinding:
sa: tom <-> cr:cr-ns

Operator

自研：

- Redis Sentinel Cluster Operator（自研，开源）
- Redis Cluster Operator（自研，开源）
- Grafana Operator（自研，开源）
- 灰度发布 Operator（自研，基于 Istio 提供灰度发布功能）
- Kunservice Operator（自研，提供双栈 Service 功能）
- BGP 相关 Operator（自研，实现网络功能）

开源：

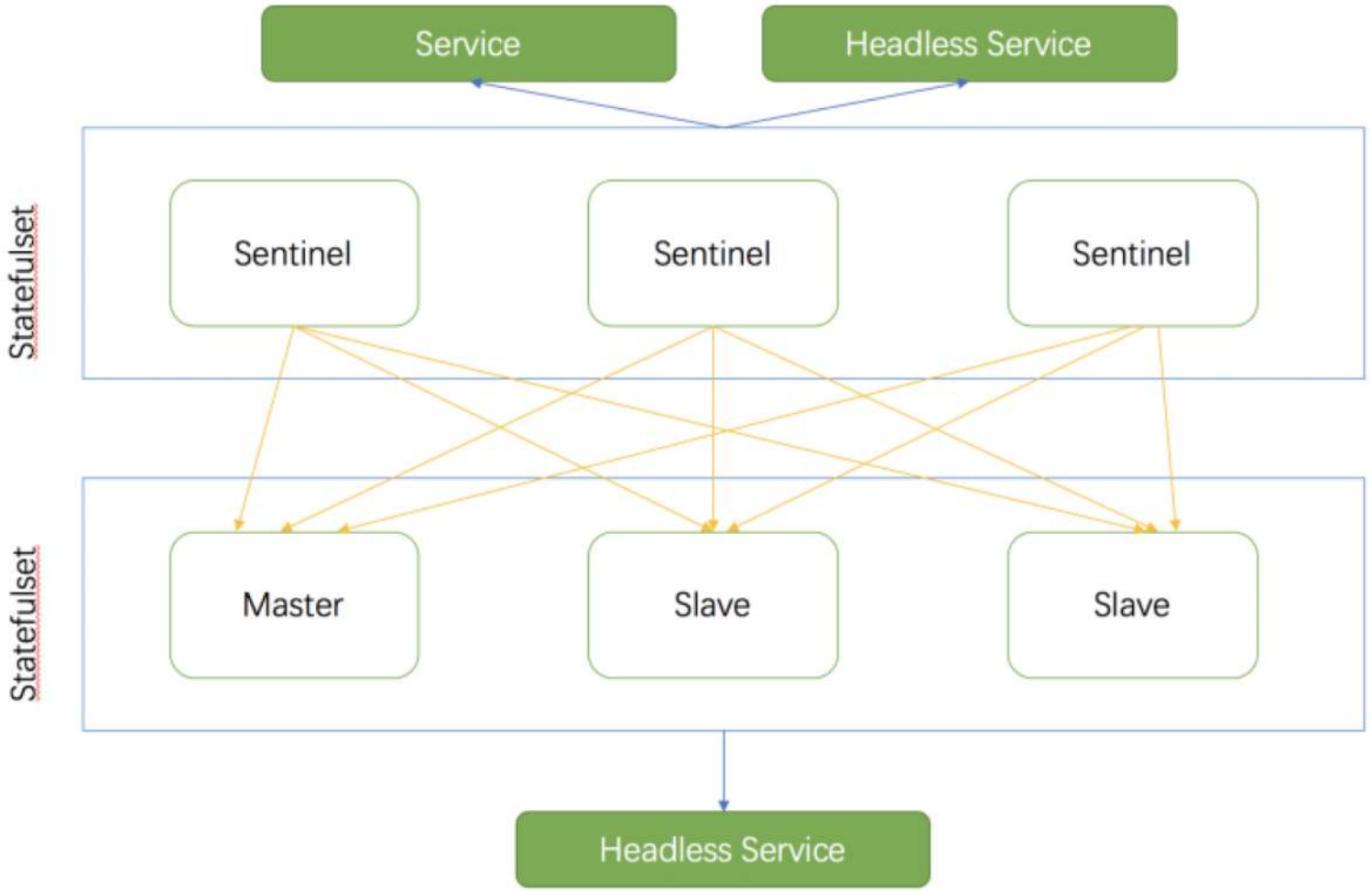
- Prometheus Operator
- Etcd Operator
- Argo Workflow, Argo Events
- NATS Operator

Redis Operator（Sentinel）

https://github.com/ucloud/redis-operator

特性

- 实时响应集群的配置变化
- 动态扩缩容
- 自愈

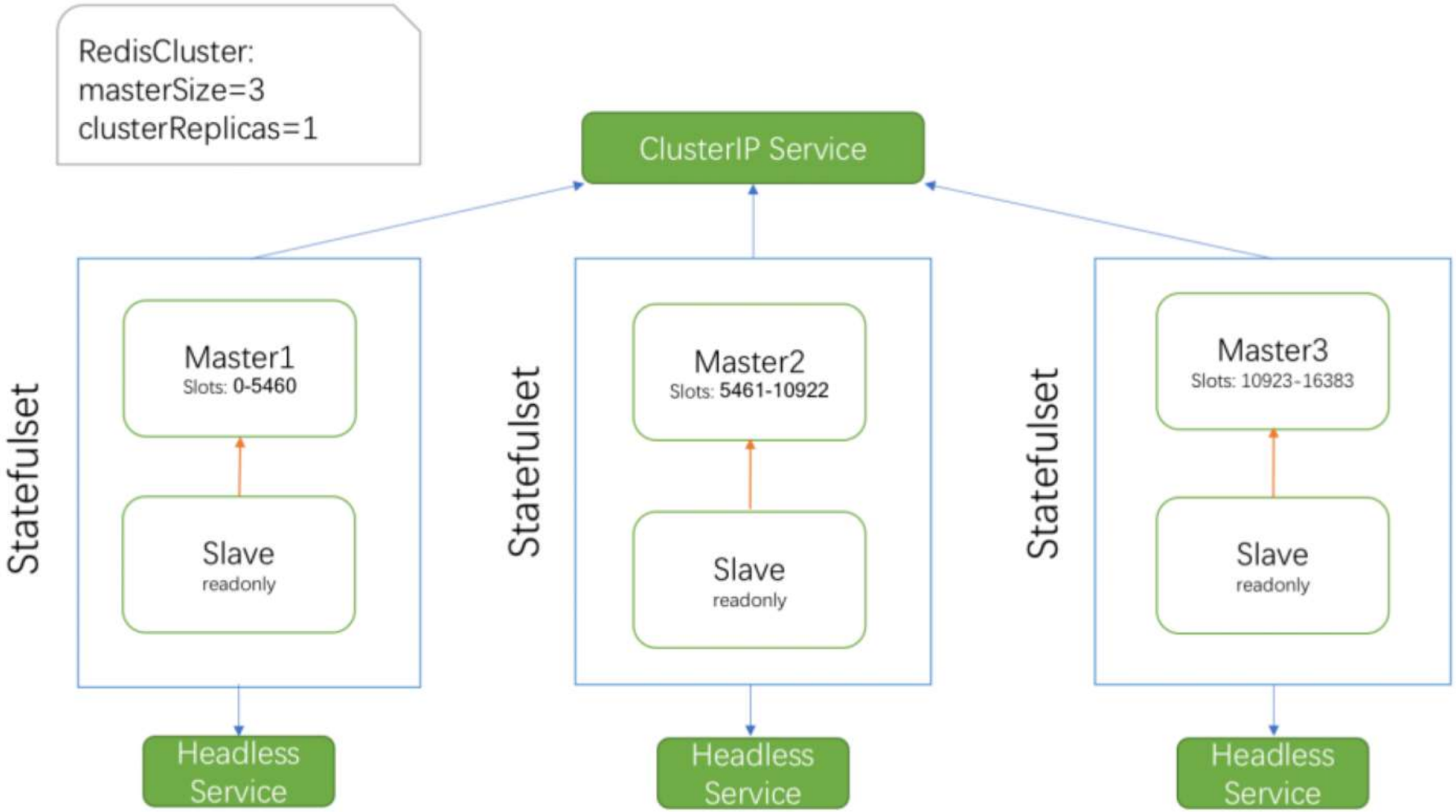


Redis Cluster Operator

https://github.com/ucloud/redis-cluster-operator

Features

- Customize the number of master nodes and the number of replica nodes per master
- Password
- Safely Scaling the Redis Cluster
- Backup and Restore
- Persistent Volume
- Custom Configuration
- Prometheus Discovery

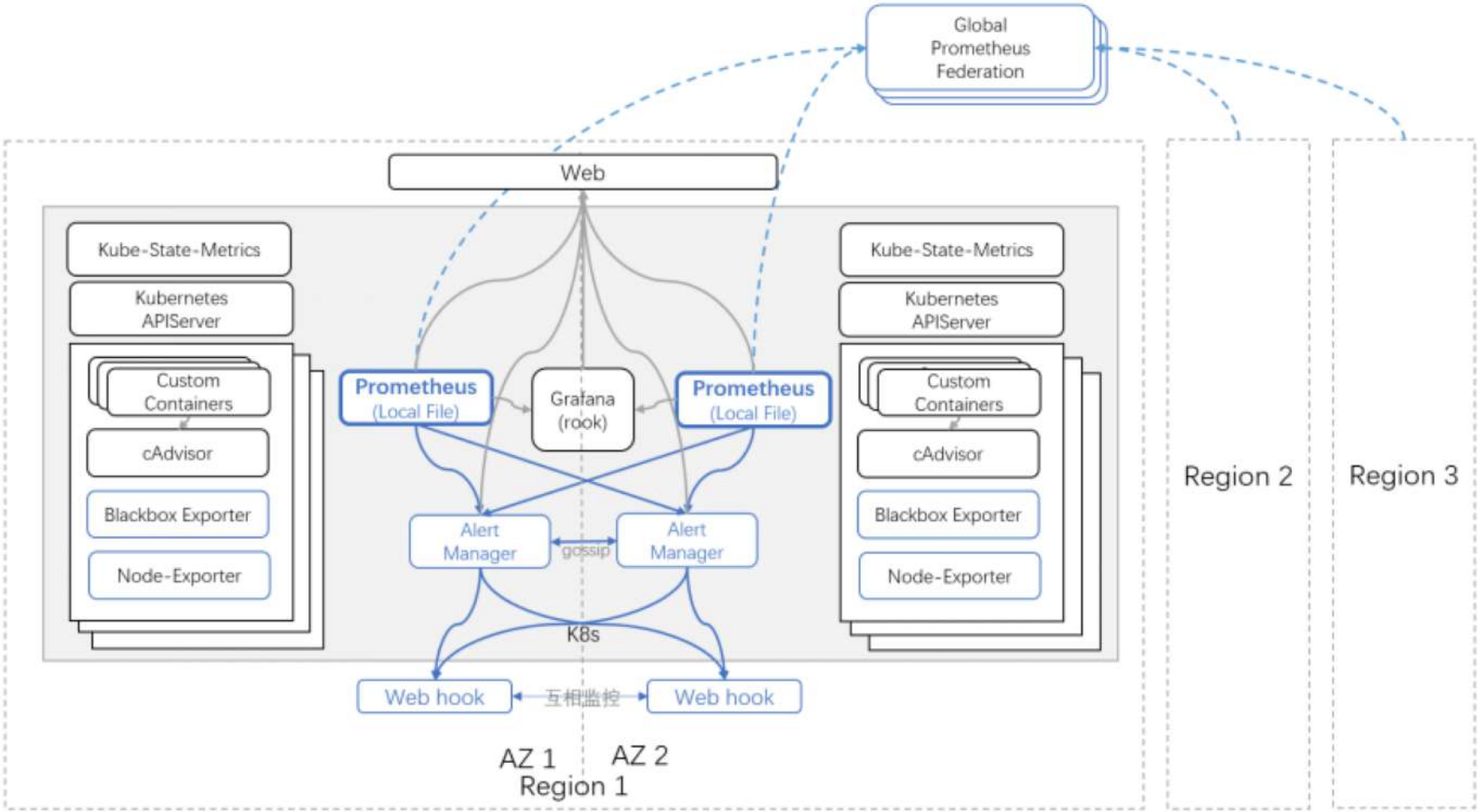


CI/CD

- Gitlab + Kubernetes
- 改造 Gitlab Runner Kubernetes Executor 实现真正“共享”：
  - 接入 IAM 进行鉴权认证
  - 对接用户所在的 Namespace 以及用户个人权限
  - 虚拟用户
- 使用 Kaniko 进行 Docker 镜像构建
- 研发“部署系统”，提供 API，IAM 鉴权，方便用户进行部署

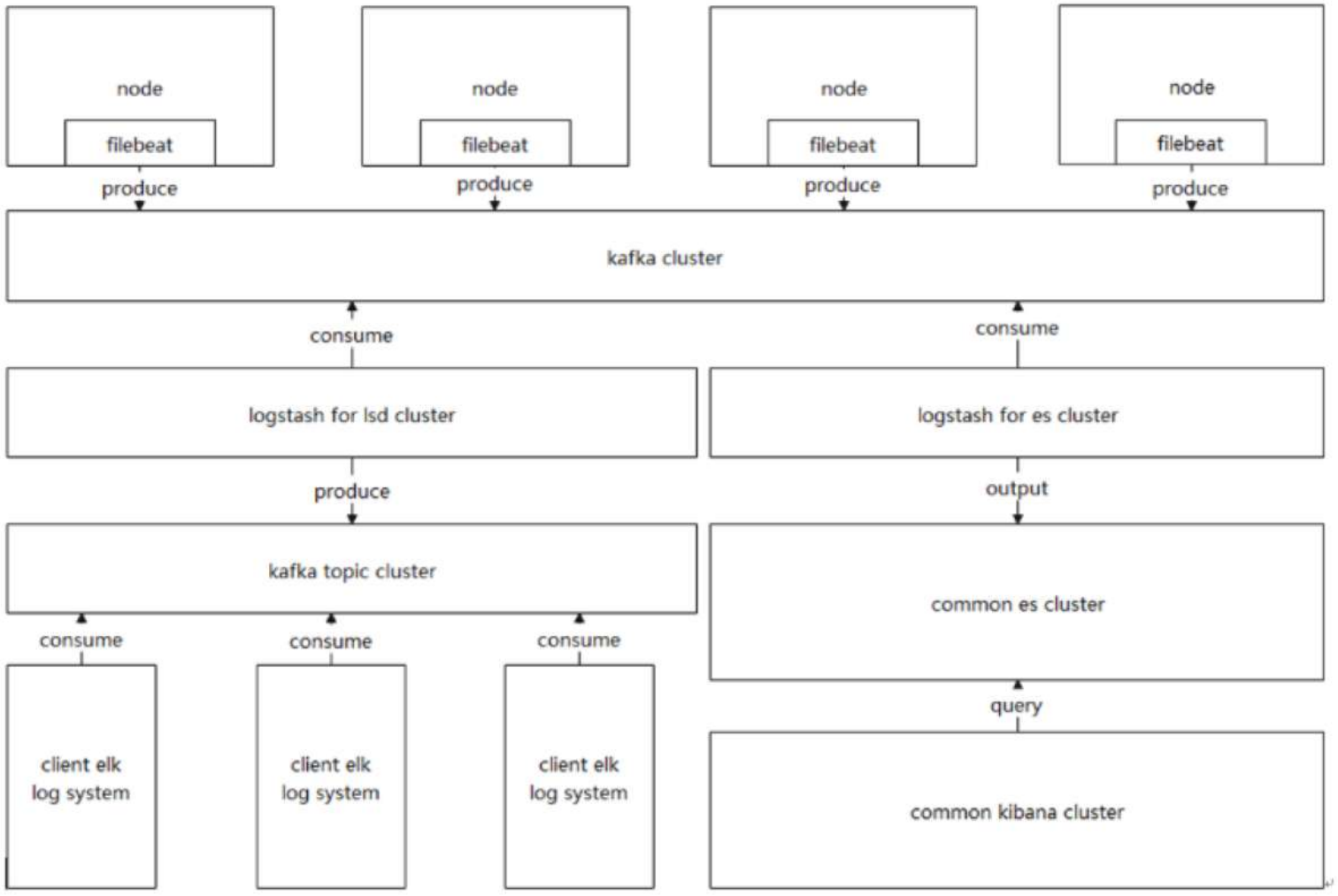
监控

- **公共监控**
  - Prometheus(HostPath Storage) 冗余
  - Exporter（自研 + 开源）
  - Alertmanager 冗余 + Webhook + NOC（微信、企业微信、邮件、人工）
  - Annotation 拨测开关
  - Webhook 部署于 k8s 外部
  - 外部商业拨测监控
- **业务定制监控**
  - Prometheus Operator
  - NOC Webhook API
  - Grafana Operator



日志

- Filebeat + Kafka + Logstash + ES
- 每个 Node 部署 Filebeat
- 两套 Kafka
  - 公共 Kafka
  - 以 Namespace 分 Topic Kafka，用于给用户消费，自己解析日志



Istio

- 基于 Istio 提供灰度发布功能
  - 提供更简单的 CRD
  - Web 操作界面
- 对 Istio 早期版本(v1.0, v1.1) 进行改造：
  - IPv6 支持
  - Namespace 隔离
  - 裁剪 Mixer Check 功能

部署

- 一个地域 1-2 个集群，多租户共享
- Ansible 部署
- 物理机为主
- Master
  - Systemd 管理 Kubelet（standalone）、Docker
  - StaticPod 启动 etcd, apiserver, controller-manager, scheduler, scheduler-extender
- Node
  - Systemd 管理 Kubelet、Docker、gobgpd
  - IPVS 实现 apiserver 高可用

运营概况

云原生社区 Cloud Native Community
UCLOUD优刻得

- 从 v1.10 开始使用，目前版本 v1.15，v1.17
- 13000+ Pod，10000+ Service
- 70+ Redis 集群
- 目前每天在 k8s 上运行 1700+ CI/CD Job
- 60000+ Image Tag

云原生社区Meetup

第一期·上海站