阿里文娱集团 | YOUKU 酷喵 UC 土豆 阿里巴巴影业集团 Alibaba Pictures 阿里巴巴·文学 Alibaba·Literature 阿里游戏 淘票票 大麦 虾米音乐 XIAMI MUSIC

UC基于Kubernetes混合云的选型和架构

张瑜标(花名:龙轼) 阿里巴巴技术专家
前京东Hadoop负责人,Hadoop代码贡献者
目前在UC基础架构和运维部
负责UC 基于Kubernetes自研的PaaS平台整体稳定性
主要专注于Service Discovery和Observability方面

目录 CONTENTS

混合云的状况和发展

稳定

稳定是一切的前提, Kubernetes 高可用集群， SLA 99.99%以上

成本

即付即用、容器化提升利用率

UAE 2.0
混合云

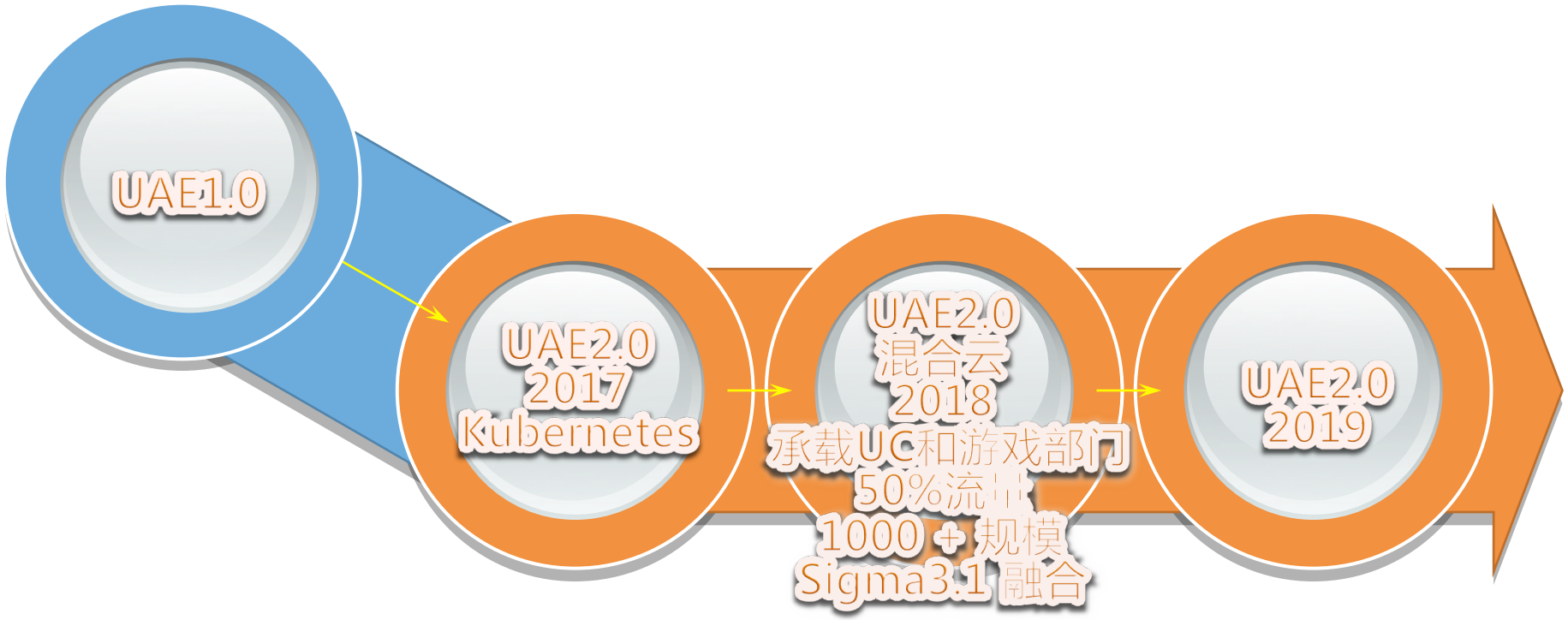
效率

系统自动化、快速部署伸缩容、赋能

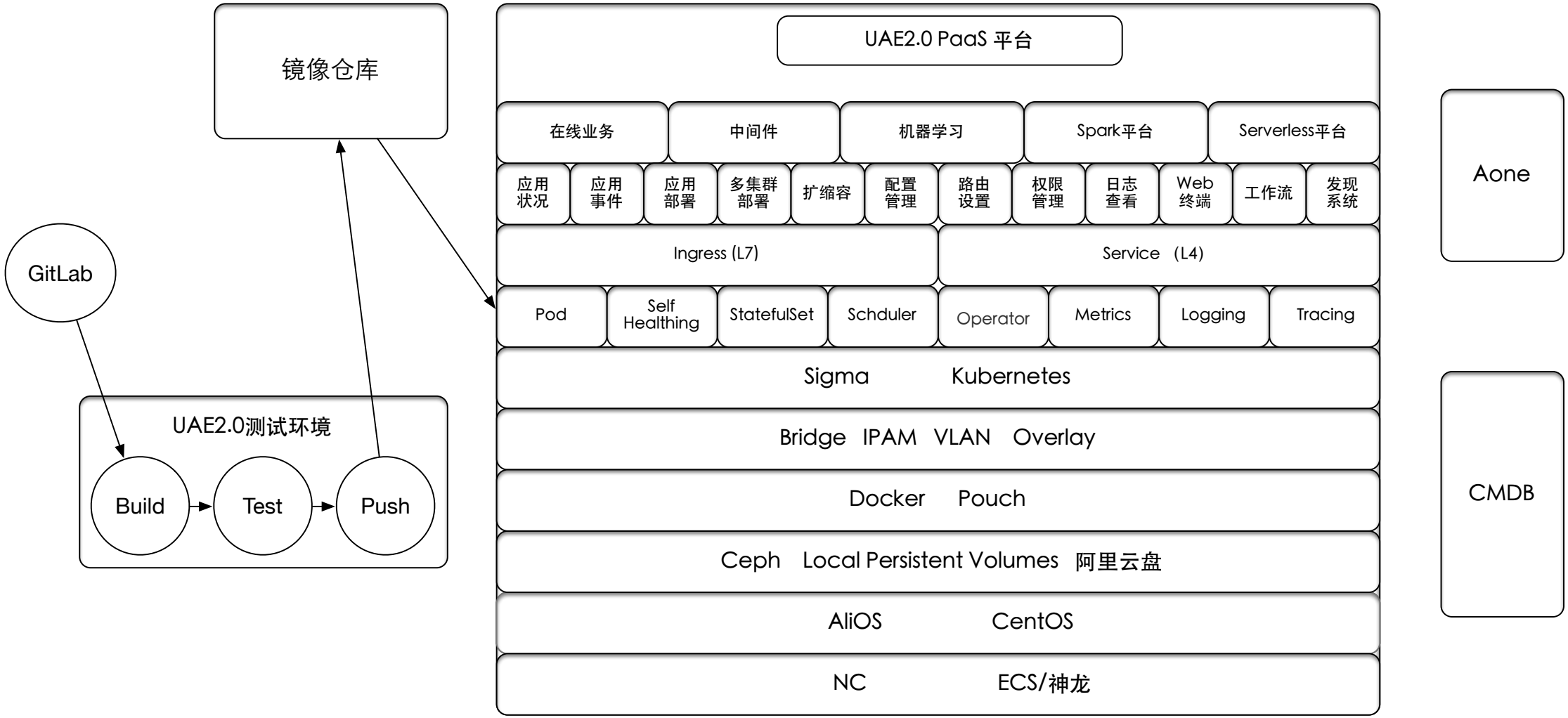
安全

私有云数据安全性

混合云的状况和发展

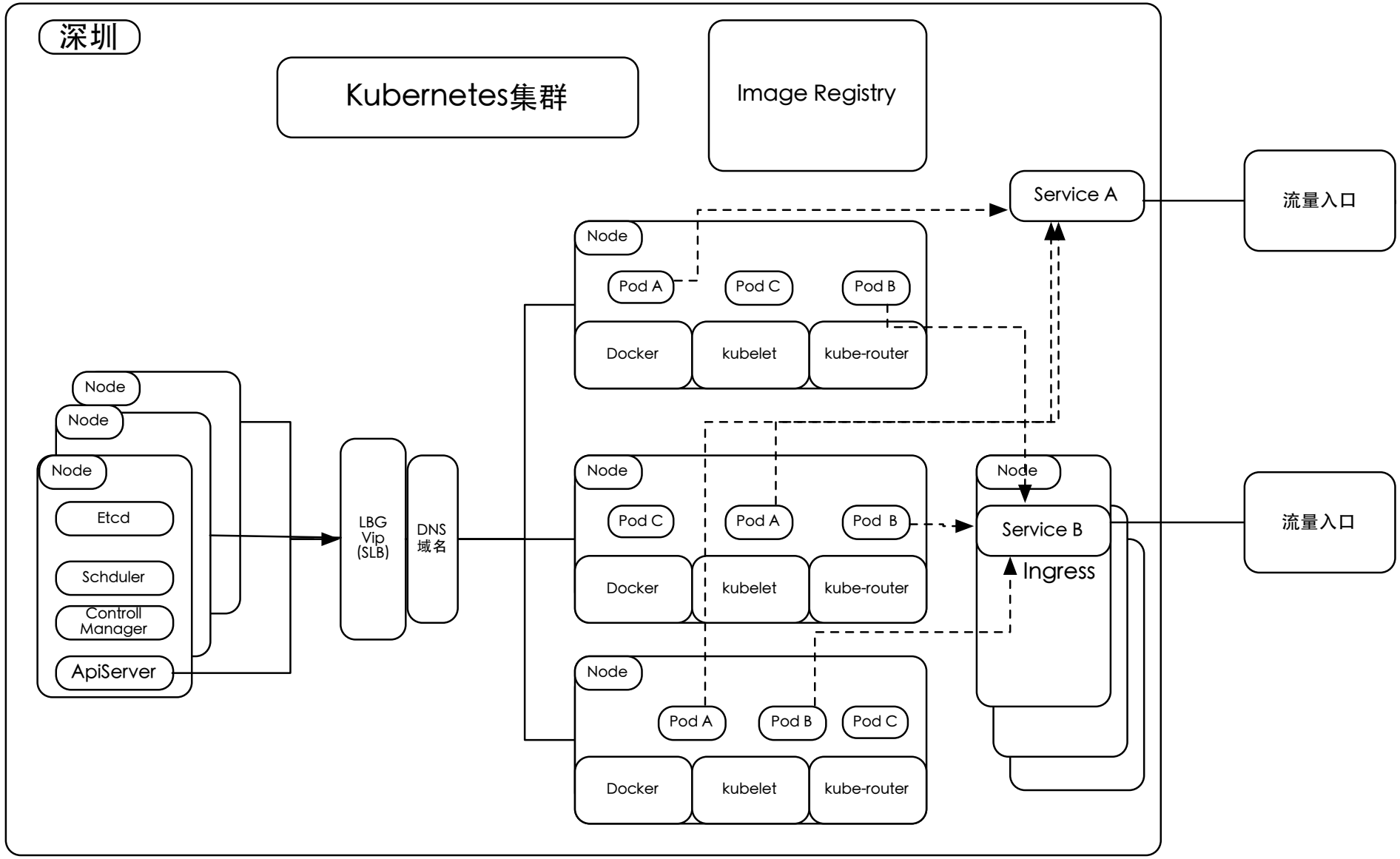

UAE1.0
UAE2.0
2017
Kubernetes
UAE2.0
混合云
2018
承载UC和游戏部门
50%流量
1000+规模
Sigma3.1融合
UAE2.0
2019

混合云的选型和架构



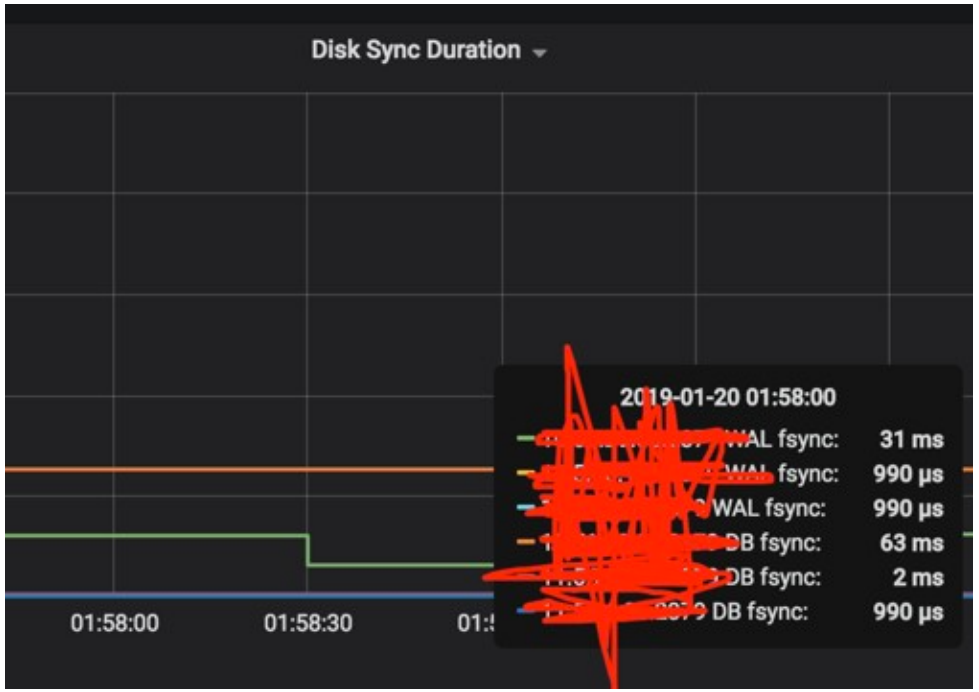
https://kubernetes.io/blog/2018/04/13/local-persistent-volumes-beta/

混合云的选型和架构


深圳
Kubernetes集群
Image Registry
Service A
流量入口
Node
Pod A
Pod C
Pod B
Docker
kubelet
kube-router
Node
Node
Node
Etcd
Schduler
Controll Manager
ApiServer
LBG Vip (SLB)
DNS 域名
Node
Pod C
Pod A
Pod B
Docker
kubelet
kube-router
Node
Service B
Ingress
流量入口
Node
Pod A
Pod B
Pod C
Docker
kubelet
kube-router

混合云的选型和架构

高可用	机柜高可用 组件高可用 SLB高可用
一体化	泛域名证书 *.uc.cn hyperkube镜像 static pod
性能	SSD VS SATA

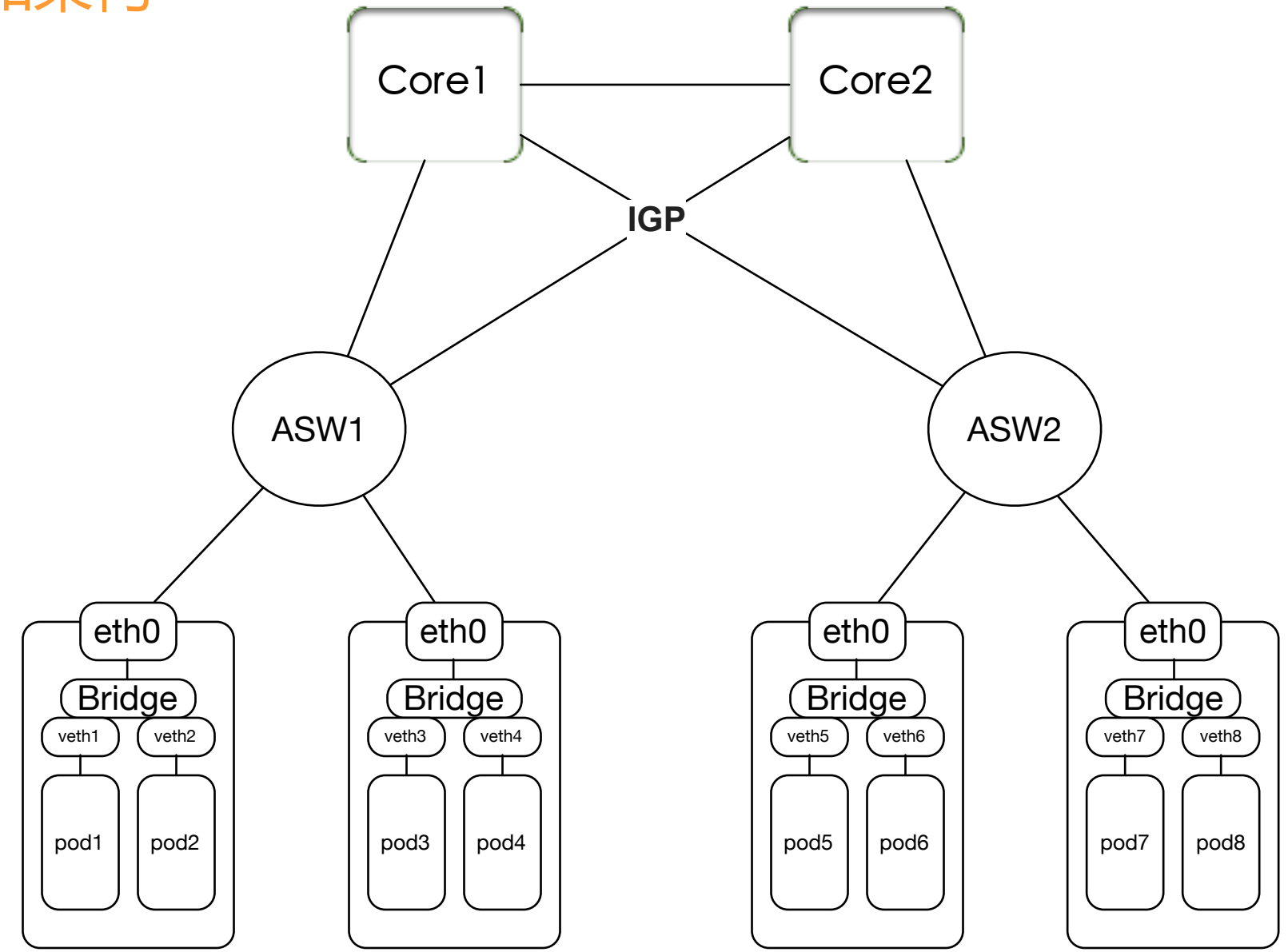


https://kubernetes.io/zh/docs/tasks/administer-cluster/static-pod/

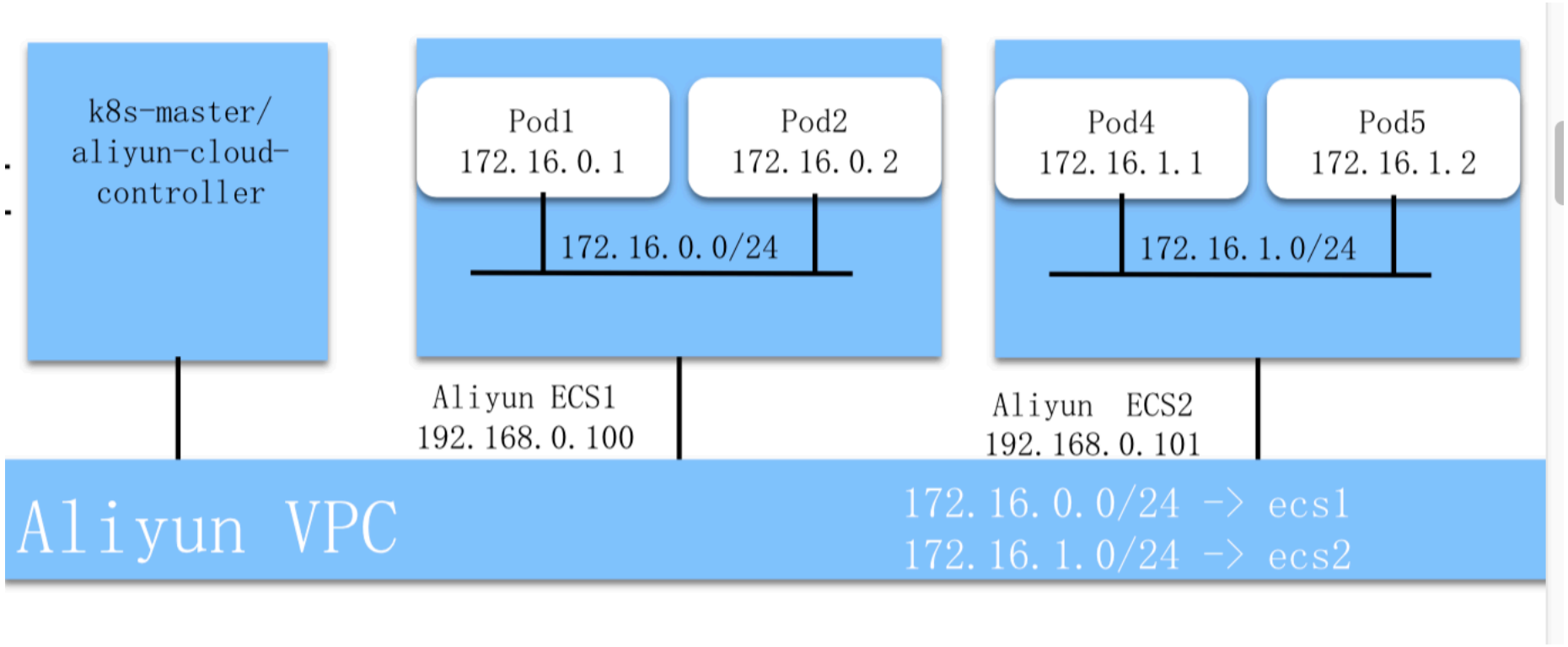
混合云的选型和架构


1
容器网络
UC：自有CNI插件
阿里云: Terway
2
持久化存储
UC : Local Persistent Volumes 和 Ceph
阿里云： 阿里云云盘
3
Service网络
UC： kube-router (ipvs) -> istio
阿里云： kube-proxy (1.12 支持ipvs)

混合云的选型和架构



混合云的选型和架构


k8s-master/
aliyun-cloud-
controller
Pod1
172.16.0.1
Pod2
172.16.0.2
172.16.0.0/24
Pod4
172.16.1.1
Pod5
172.16.1.2
172.16.1.0/24
Aliyun ECS1
192.168.0.100
Aliyun ECS2
192.168.0.101
Aliyun VPC
172.16.0.0/24 -> ecs1
172.16.1.0/24 -> ecs2

混合云的选型和架构

1

Kubernetes
惊天地泣鬼神
之大Bug

Endpoints问题

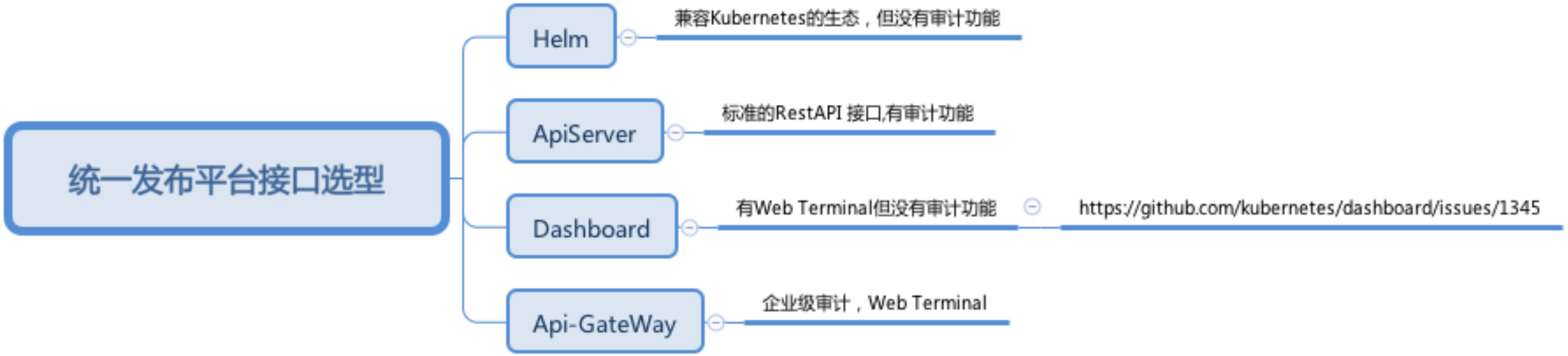
2

kubelet 节点du
导致iowait问题

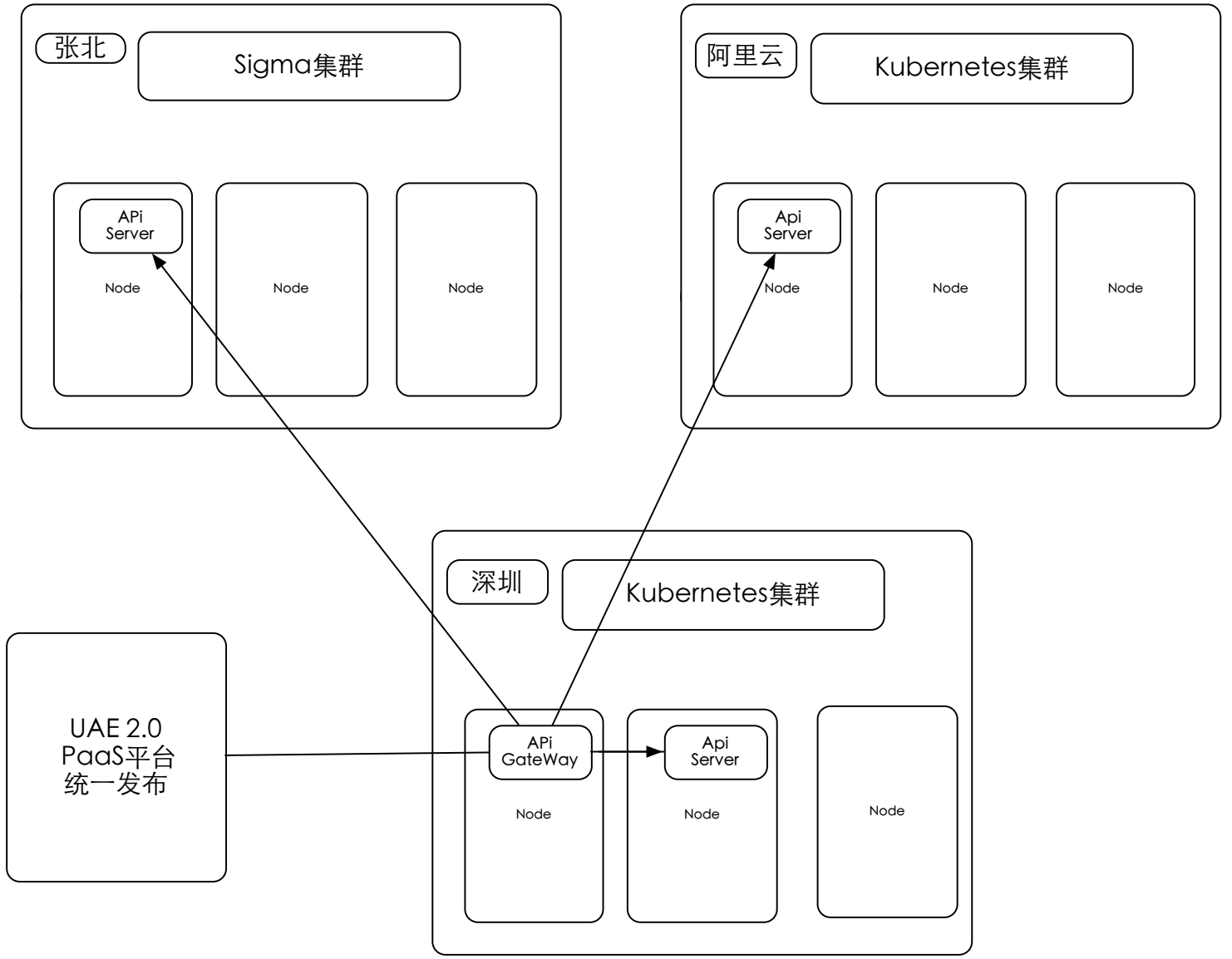
iowait问题

https://zhuanlan.zhihu.com/p/37217575
https://github.com/kubernetes/kubernetes/pull/58547
https://github.com/kubernetes/kubernetes/issues/58217
https://github.com/kubernetes/kubernetes/issues/61999

统一发布系统


统一发布平台接口选型
Helm
兼容Kubernetes的生态，但没有审计功能
ApiServer
标准的RestAPI 接口,有审计功能
Dashboard
有Web Terminal但没有审计功能
https://github.com/kubernetes/dashboard/issues/1345
Api-GateWay
企业级审计，Web Terminal

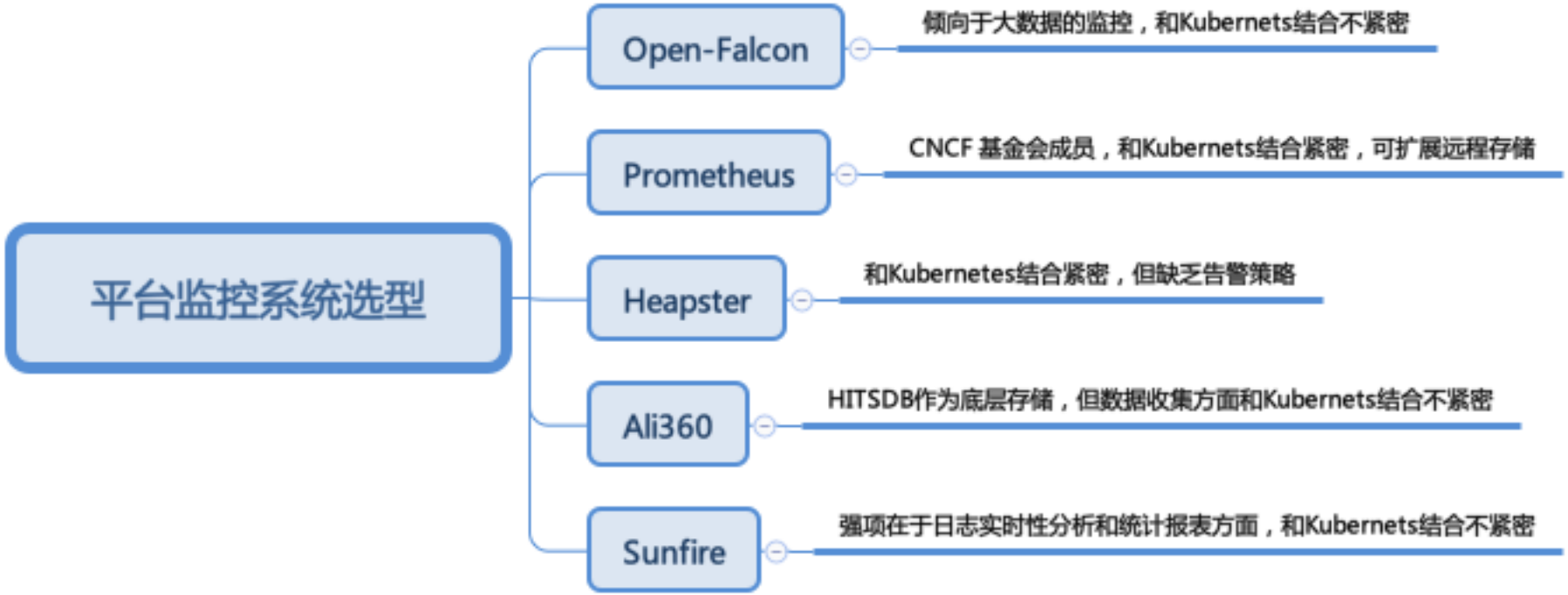
统一发布系统


张北
Sigma集群
APi Server
Node
Node
Node
阿里云
Kubernetes集群
Api Server
Node
Node
Node
深圳
Kubernetes集群
UAE 2.0
PaaS平台
统一发布
APi GateWay
Node
Api Server
Node
Node

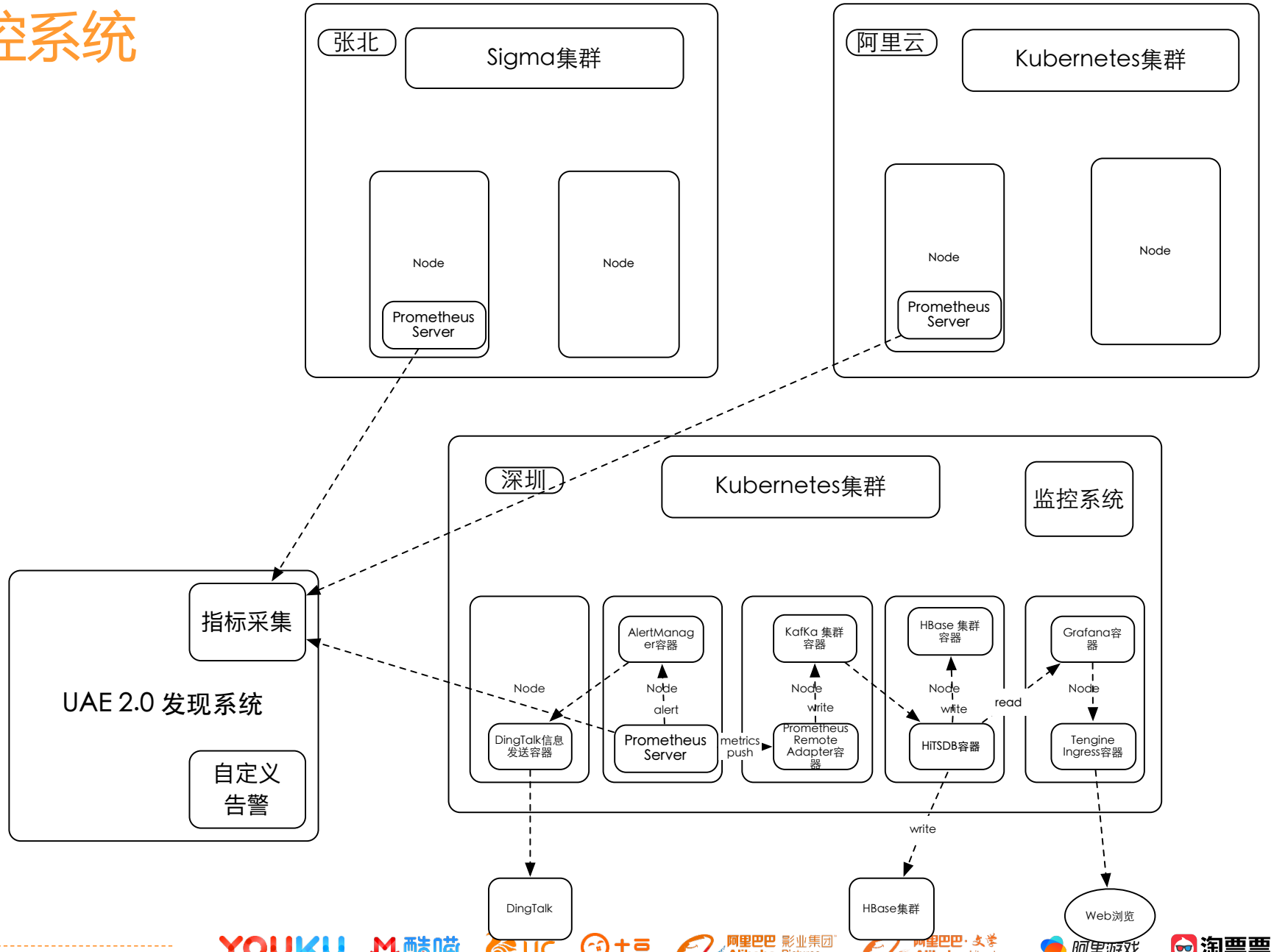
统一发布系统

审计	企业级审计 审计日志上传
Web登录	Websocket WebTerminal
多集群	多集群支持 正在兼容helm

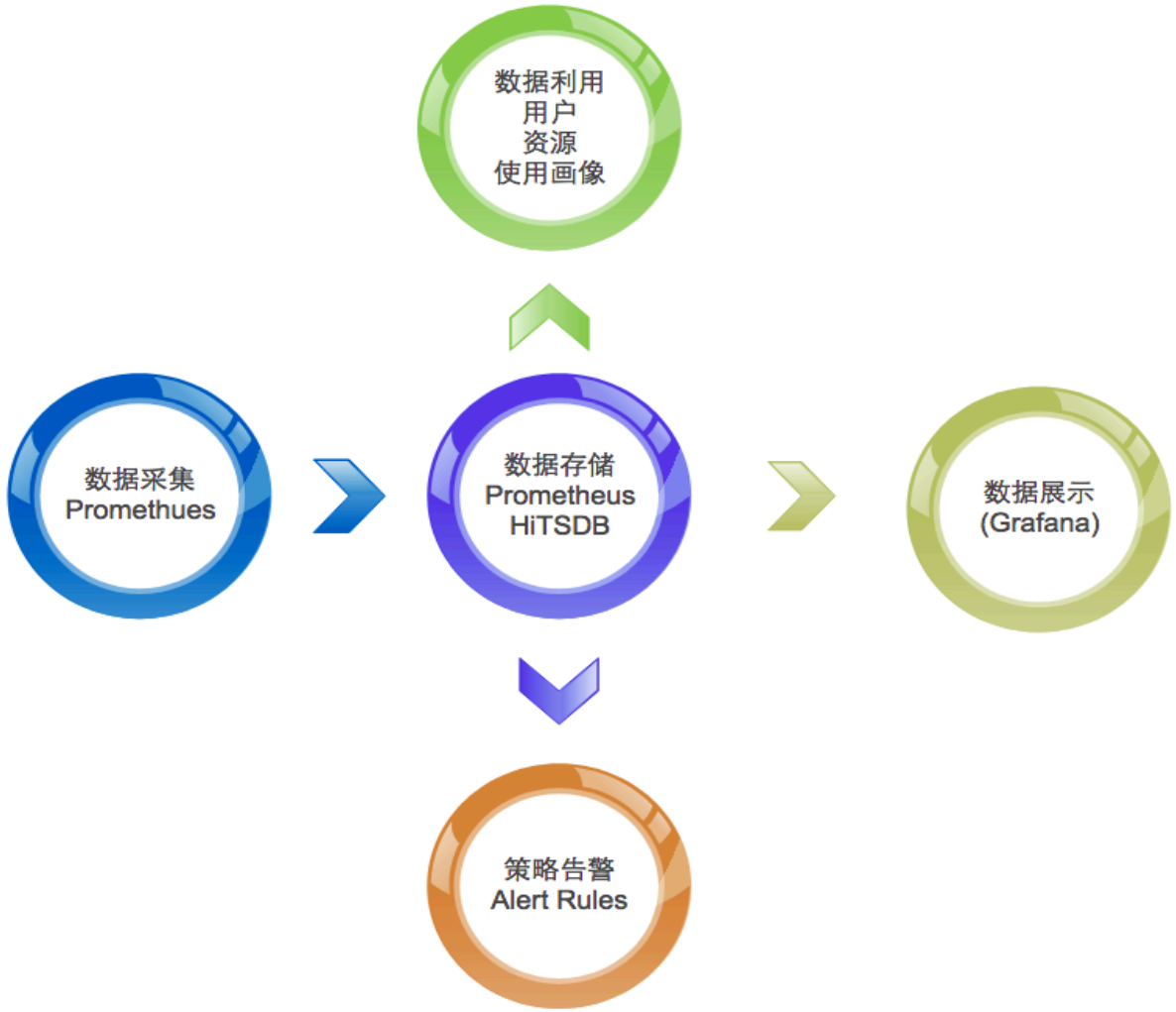
统一监控系统


平台监控系统选型
Open-Falcon
倾向于大数据的监控，和Kubernets结合不紧密
Prometheus
CNCF 基金会成员，和Kubernets结合紧密，可扩展远程存储
Heapster
和Kubernetes结合紧密，但缺乏告警策略
Ali360
HITSDB作为底层存储，但数据收集方面和Kubernets结合不紧密
Sunfire
强项在于日志实时性分析和统计报表方面，和Kubernets结合不紧密

统一监控系统


张北
Sigma集群
Node
Prometheus Server
Node
阿里云
Kubernetes集群
Node
Prometheus Server
Node
深圳
Kubernetes集群
监控系统
指标采集
UAE 2.0 发现系统
自定义告警
Node
DingTalk信息发送容器
AlertManager容器
Node
alert
Prometheus Server
metrics push
KafKa 集群容器
Node
write
Prometheus Remote Adapter容器
HBase 集群容器
Node
write
HiTSDB容器
read
Grafana容器
Node
Tengine Ingress容器
write
DingTalk
HBase集群
Web浏览

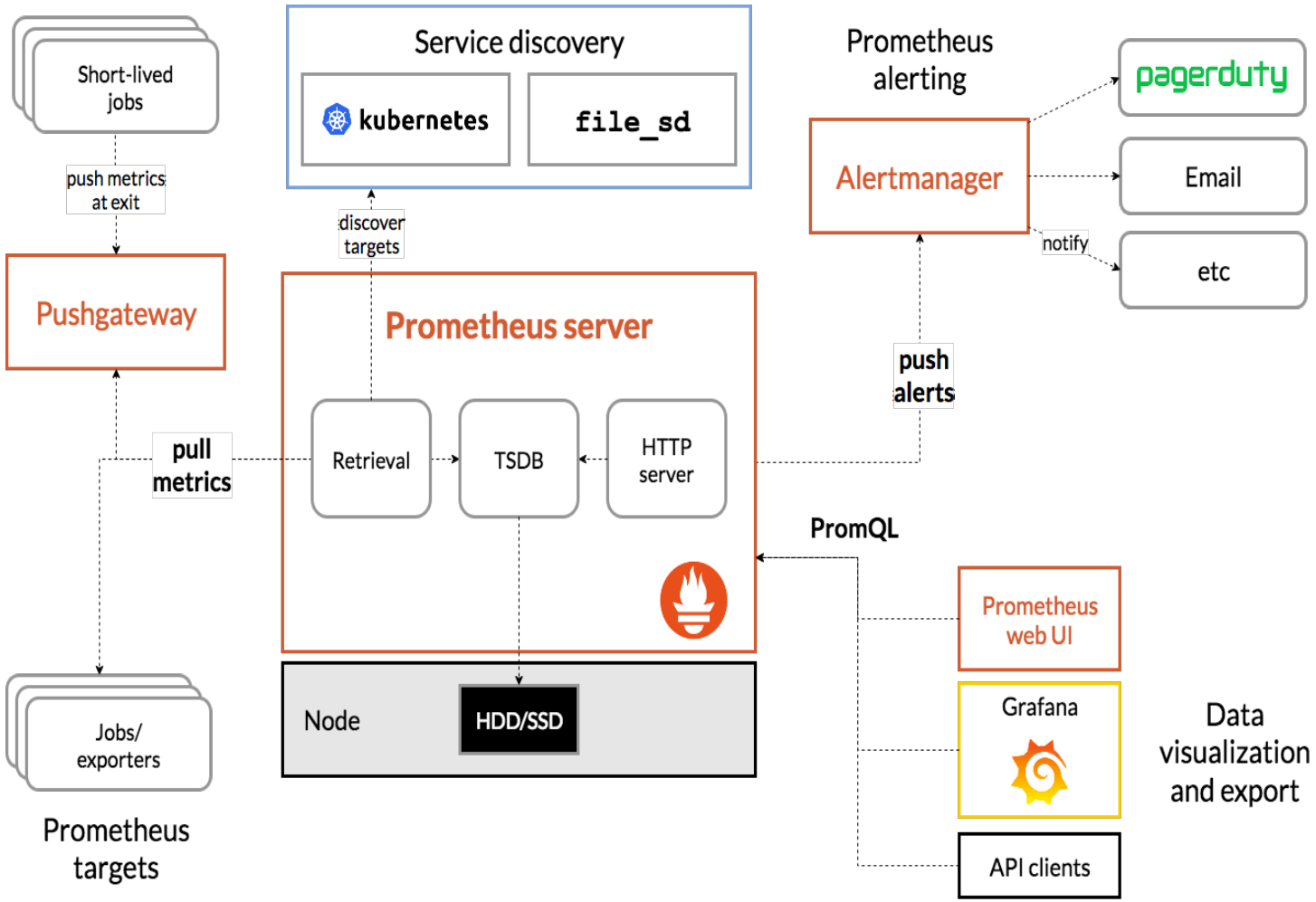
统一监控系统


数据利用
用户
资源
使用画像
数据采集
Promethues
数据存储
Prometheus
HiTSDB
数据展示
(Grafana)
策略告警
Alert Rules

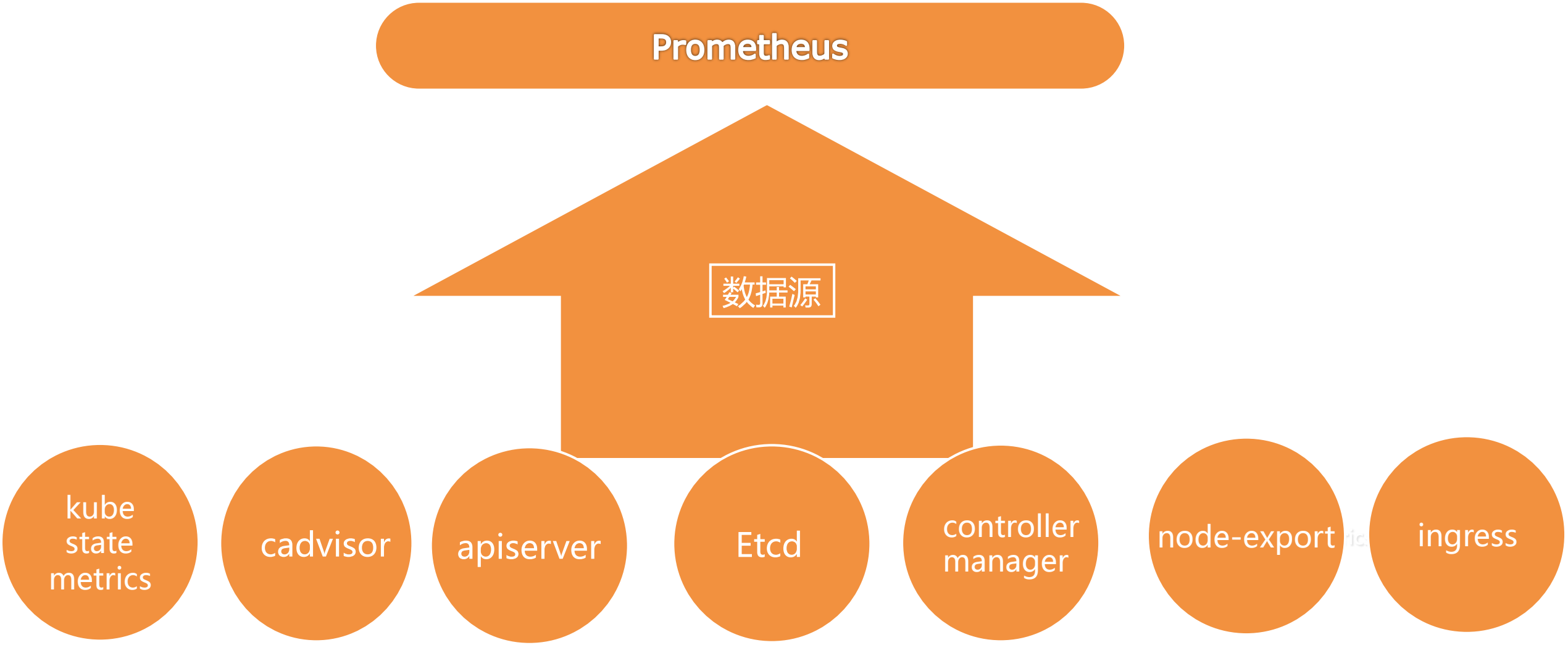
统一监控系统

高可用	机柜考量 组件考量
兜底	外部探测系统UCMT
性能	SSD 本地短时存储 中转到分布式时序数据库

统一监控系统

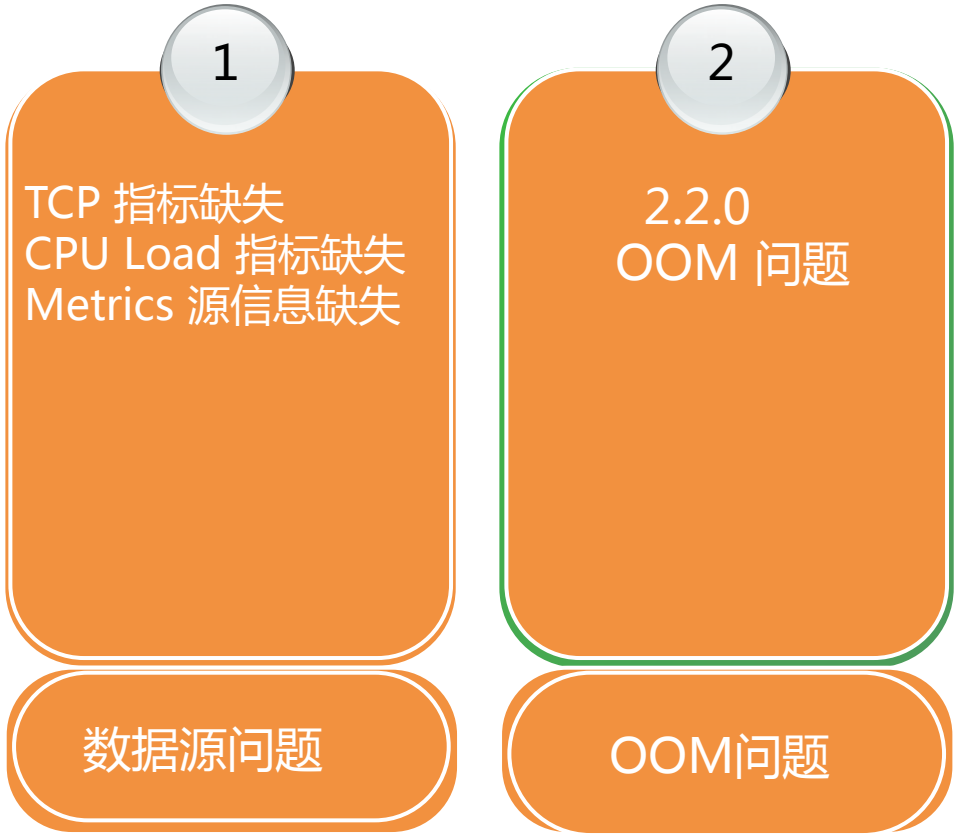

Short-lived jobs
push metrics at exit
Pushgateway
Service discovery
kubernetes
file_sd
discover targets
Prometheus server
Retrieval
TSDB
HTTP server
pull metrics
Node
HDD/SSD
Jobs/ exporters
Prometheus targets
Prometheus alerting
Alertmanager
pagerduty
Email
etc
notify
push alerts
PromQL
Prometheus web UI
Grafana
API clients
Data visualization and export

统一监控系统


Prometheus
数据源
kube state metrics
cadvisor
apiserver
Etcd
controller manager
node-export
ingress

统一监控系统

统一监控系统



https://stackoverflow.com/questions/49083348/cadvisor-prometheus-integration-returns-container-cpu-load-average-10s-as-0

https://github.com/google/cadvisor/issues/1932

https://github.com/google/cadvisor/issues/1298

https://www.slideshare.net/BartomiejPotka/thanos-global-durable-prometheus-monitoring

统一安全方案

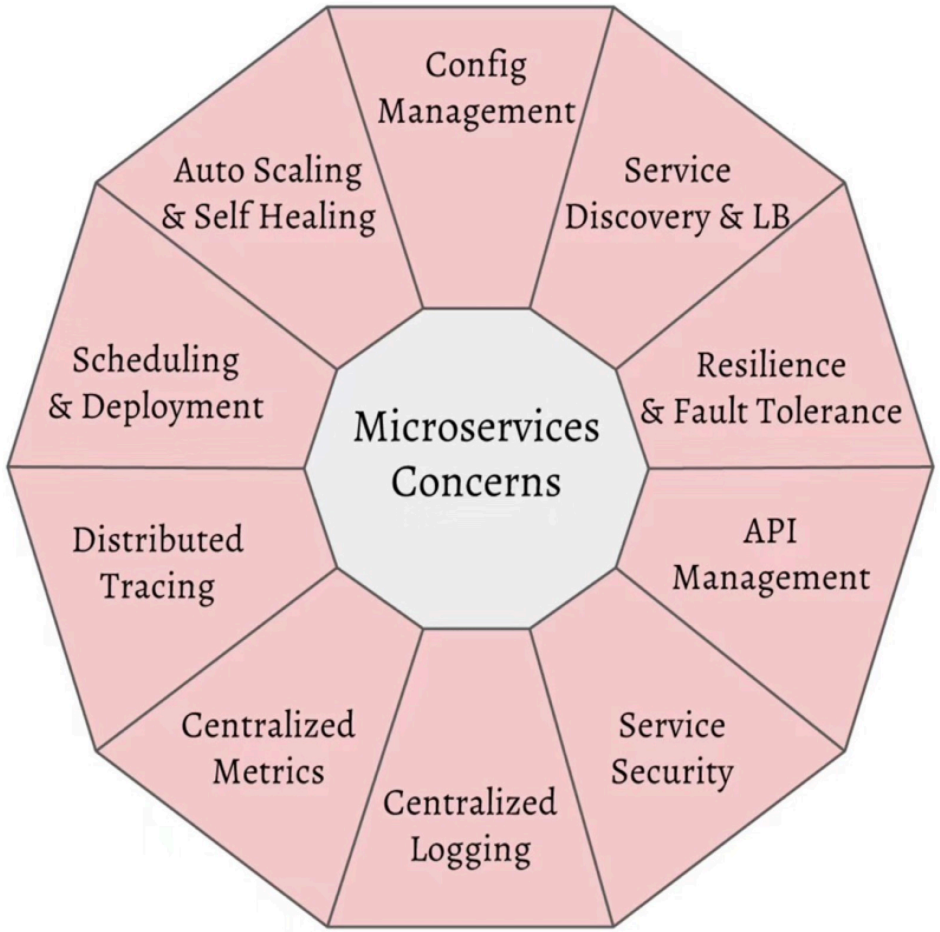
基础层	Linux 安全 端口安全
容器层	容器安全(gvisor, kata) Kubernetes 安全(RBAC)
关注	CIS

七牛云容器安全基线
https://blog.qiniu.com/archives/7743

统一服务发现方案

DNS	CoreDNS 插件模式兼容其他服务发现
Cache	Per Node 的 DNS Cache
关注	异地机房容灾 智能路由调度

后续的演进方向

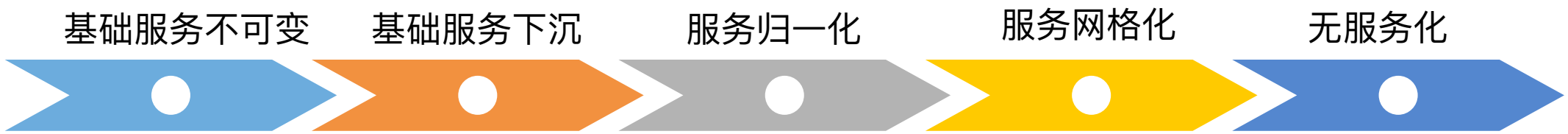

Config Management
Service Discovery & LB
Resilience & Fault Tolerance
API Management
Service Security
Centralized Logging
Centralized Metrics
Distributed Tracing
Scheduling & Deployment
Auto Scaling & Self Healing
Microservices Concerns


Users and Systems (IoT) use applications developed and deployed by developers and hosted by operators
Istio
Knative
Kubernetes
Developers build and deploy apps
API
GitHub
Contributors develop and contribute code and docs to the OSS project
Operators deploy and manage Knative instances using Kubernetes API and tools
Platform Providers (such as Google Cloud Platform) provide underlying infrastructure

后续的演进方向

引用:

- 可靠的交付流水线：让团队不用担心验证和部署的环境，步骤及流程。
- 容器技术（Docker）：让团队不必过多考虑构建分发及运行环境的问题。
- Kubernetes：让团队不用过多考虑容器应用的部署、运行、扩缩容等工作。
- Sevice Mesh：让团队不用过多考虑分布式服务的通信。
- Severless：让团队不用过多考虑服务器的实体资源。



https://thenewstack.io/ebooks/kubernetes/state-of-kubernetes-ecosystem/
http://www.servicemesher.com/blog/service-mesh-meetup-shanghai-20181125/

Thanks!

阿里云智能事业群-基础产品事业部

互动娱乐事业部-游戏研发线-技术中台

UC事业部-基础架构与运维部