LINUXCON containercon CHINA 中国 CLOUDOPEN

THINK OPEN

开放性思维

Host Kubernetes Within Kubernetes

腾讯云基础PaaS团队 于广游

LF ASIA, LLC

旧方案架构总览

多Kubernetes运维难点

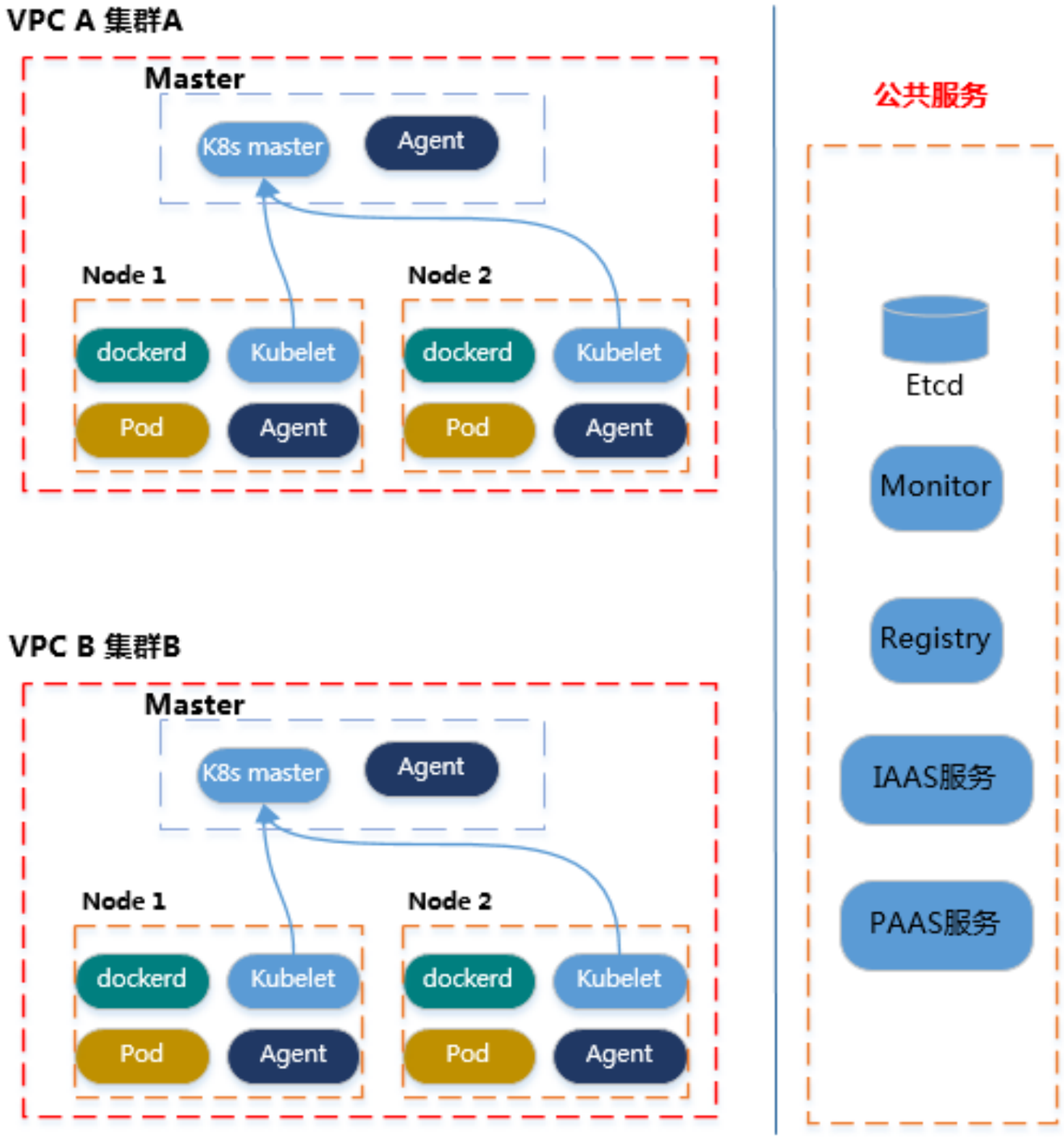
灵感来源

新方案

总结

总体架构

- **单集群用户独占**
- **网络**
  - 基于VPC实现容器网络
  - 集群在用户VPC中
- **Node**
  - CVM + Node Component
- **Master**
  - CVM + Master Component
  - 用户不可见
- **Etcd**
  - 多租户共享
- **部署**
  - CVM上部署Agent
  - Agent来初始化节点



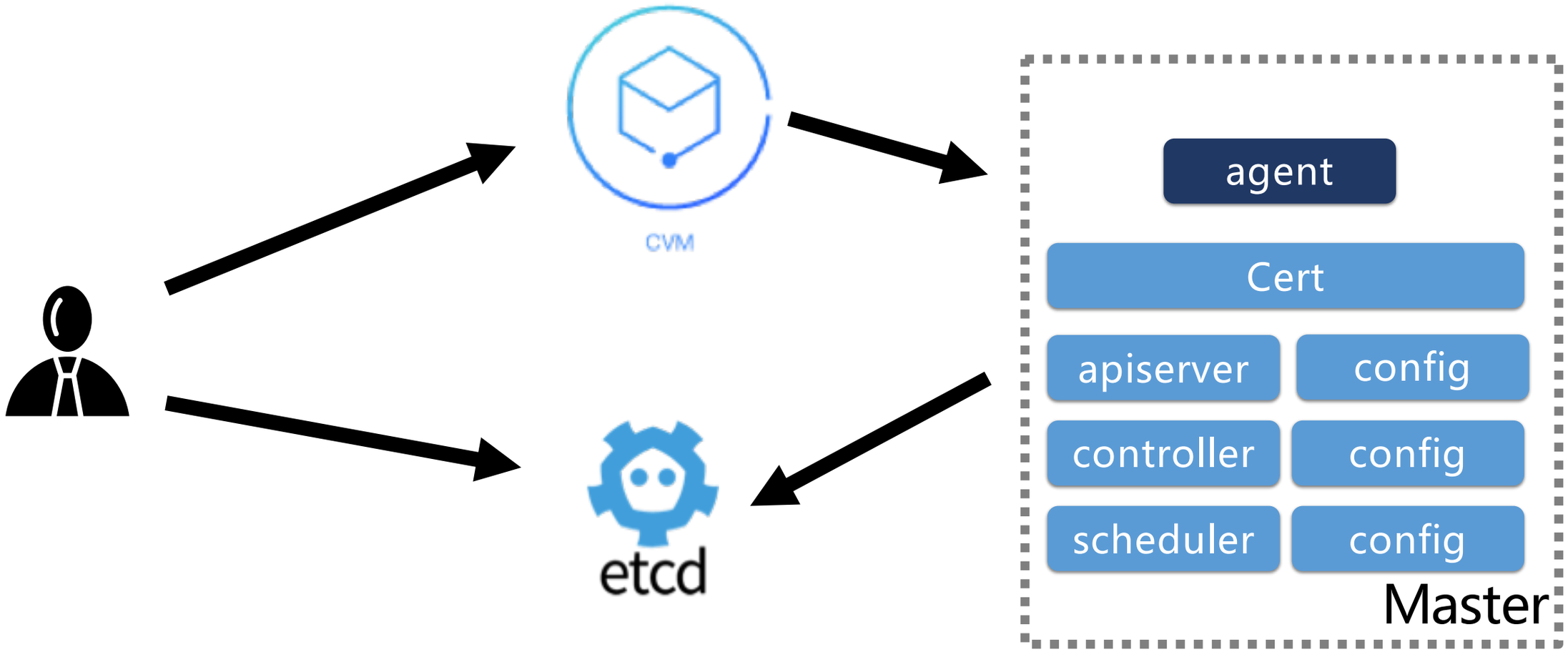
功能总览

- **集群管理**
  - 集群创建销毁
  - 集群(自动)扩缩容
- **应用编排**
  - 应用模板
  - 发布管理
- **监控服务**
  - 基础监控
  - 服务监控
  - 日志中心
- **镜像管理**
  - 镜像仓库
  - 自动构建
  - 触发器

总体架构

- **为什么Master独立CVM部署在VPC中？**
  - Master 与 Node需要双向互通
  - 不同VPC网段可能冲突
  - 简单
- **为什么Etcd共享？**
  - 部署到 Master上 -> 数据单点风险
  - 每集群一套 -> 复杂
  - 共享 -> 简单

集群创建


CVM
agent
Cert
apiserver
config
controller
config
scheduler
config
Master
etcd

旧方案架构总览

多Kubernetes运维难点

灵感来源

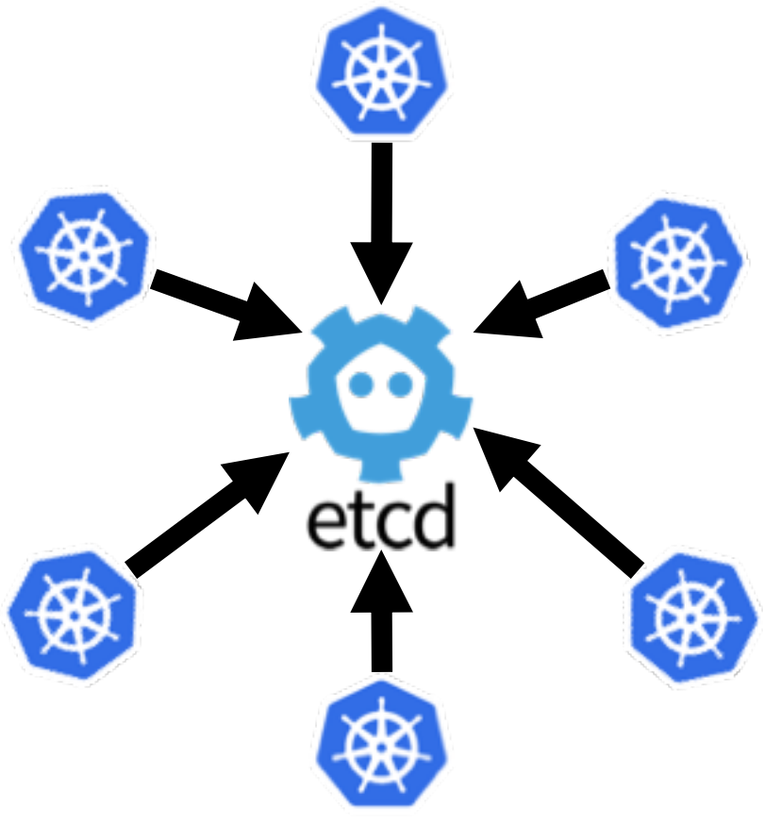
新方案

总结

旧方案非常简单
却有很多问题

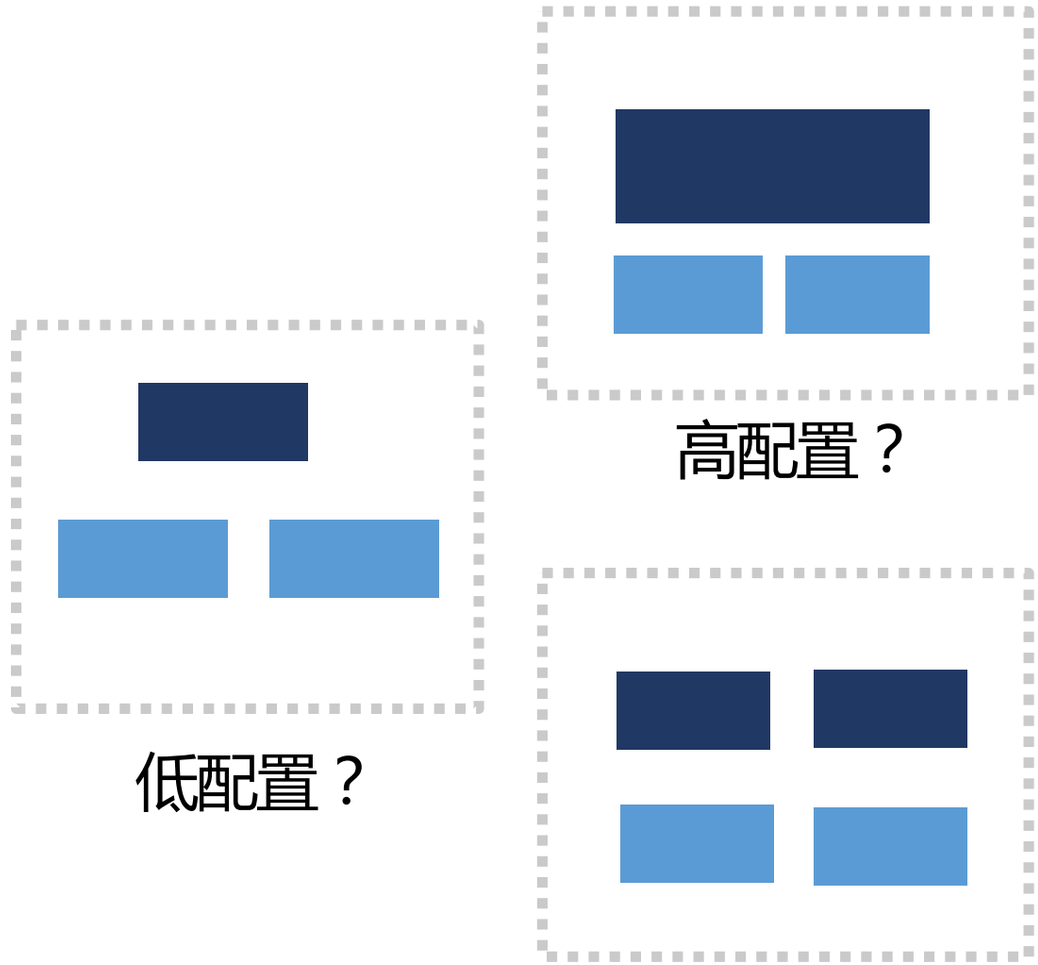
困境

- **多集群共享Etcd**
  - 性能
  - 运维
  - 可用性



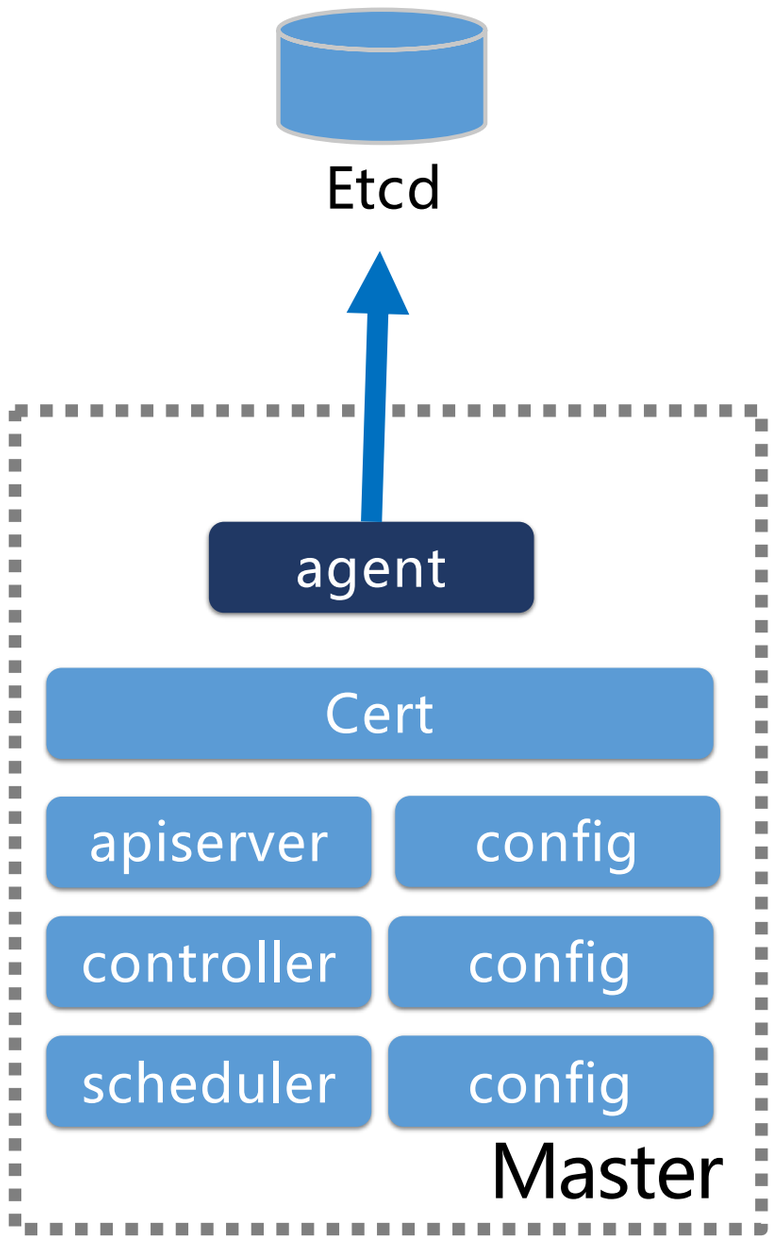
困境

- **Master为VPC内独立CVM**
  - 成本
  - 性能
  - 可用性



困境

- **agent管理节点**
  - 安装包、证书、配置(接近20个文件)
  - 后台数据与Master上实际配置一致性
  - 异常回滚
  - Agent本身设计、开发、运维



困境

- **监控**
  - 提供给用户的监控与后台使用的监控是完全不同的两套
  - 监控对象与环境不统一，维护困难
  - 监控agent发布升级难

模块	运行环境	对象	监控方式	发布方式
Etcd	支撑网络	负载 Etcd metrics Etcd log	支撑监控agent Metrics server filebeat	支撑发布
Master	用户VPC	负载 k8s metrics k8s log	CVM监控agent Metrics server filebeat	CVM监控发布 容器节点agent发布
Node	用户VPC	CVM基础监控 Pod监控	CVM监控agent cadvisor	CVM监控发布

思考

k8s为高效运维而生，为什么运维起来却这么难？

我们为用户提供了高效的k8s服务，为什么自己用不到?

用k8s来管理k8s，是不是就能解决这个问题？

旧方案架构总览

多Kubernetes运维难点

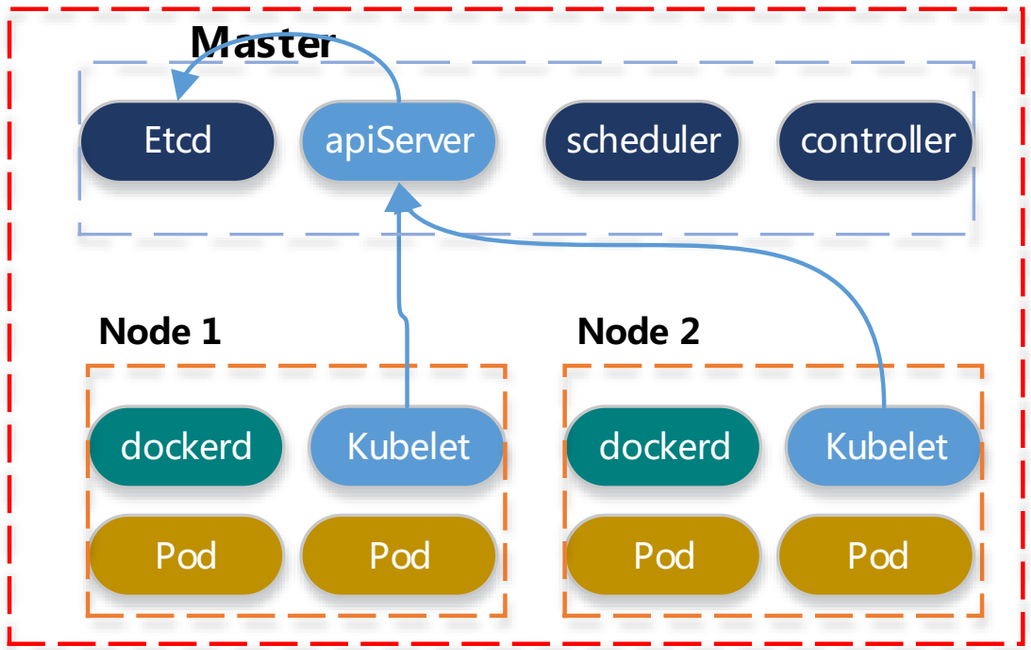
灵感来源

新方案

总结

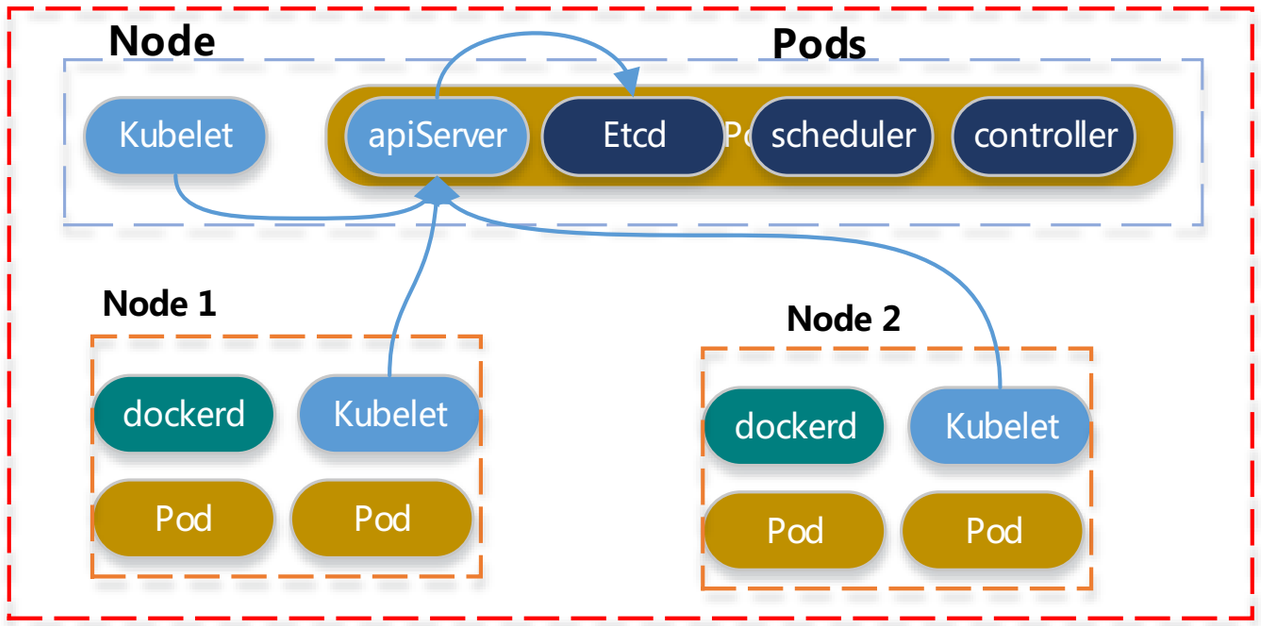
灵感

传统部署



- Master组件单独部署
- 通常SSH等管理Master组件
- Master与业务应用的运维不一致

Self-Hosted



- Master组件运行为集群中的普通Pod
- 通过k8s API来管理Master组件
- Master与业务应用的运维一致

灵感

- **Self-Hosted Kubernetes**
  - 升级
  - 健康检查
  - 扩容
  - 监控

旧方案架构总览

多Kubernetes运维难点

灵感来源

新方案

总结

新方案

目标

利用k8s来管理k8s，简化k8s的运维

问题

Master与Node通信

Etcd部署

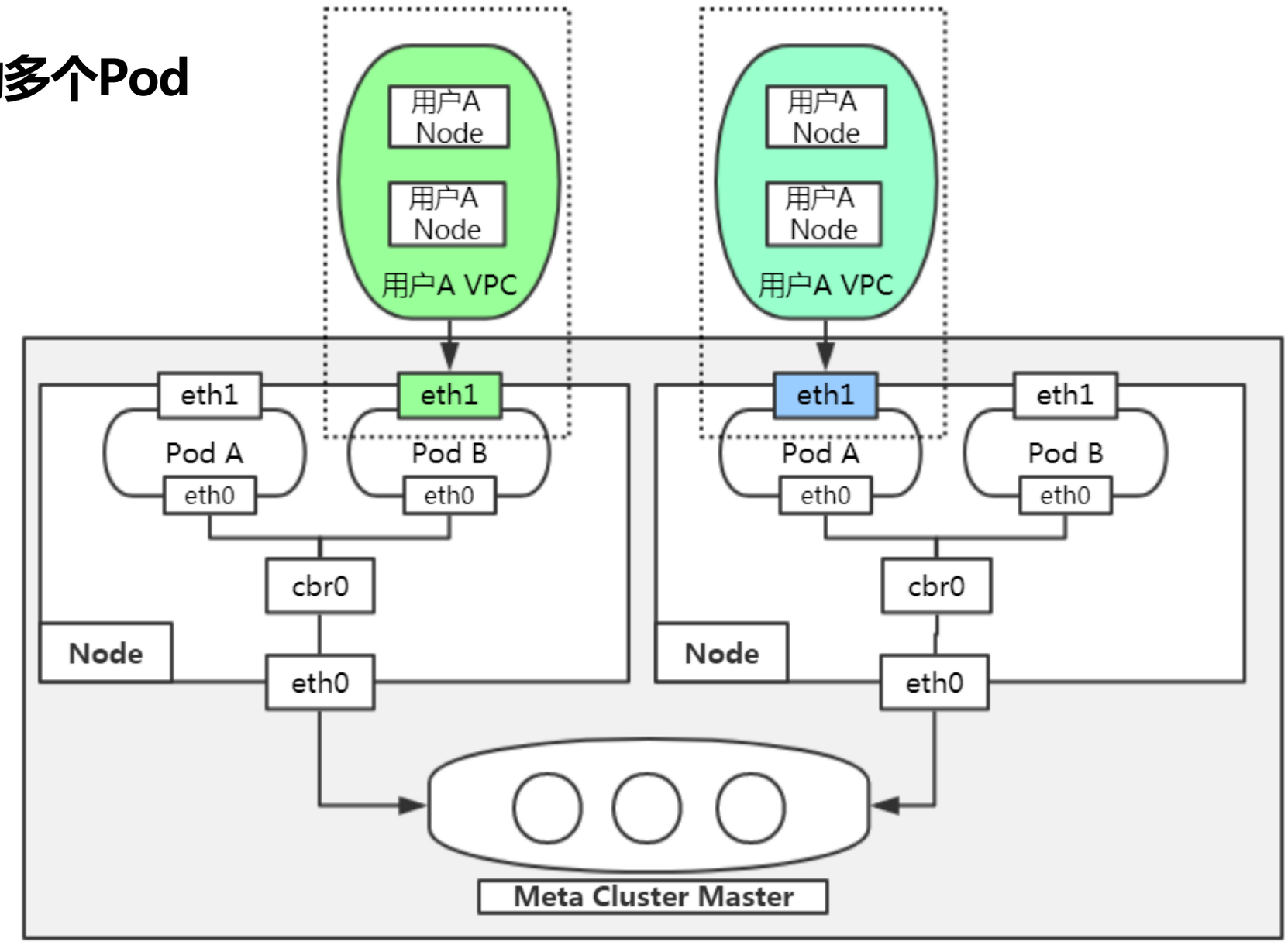
Master部署

Master组件更新

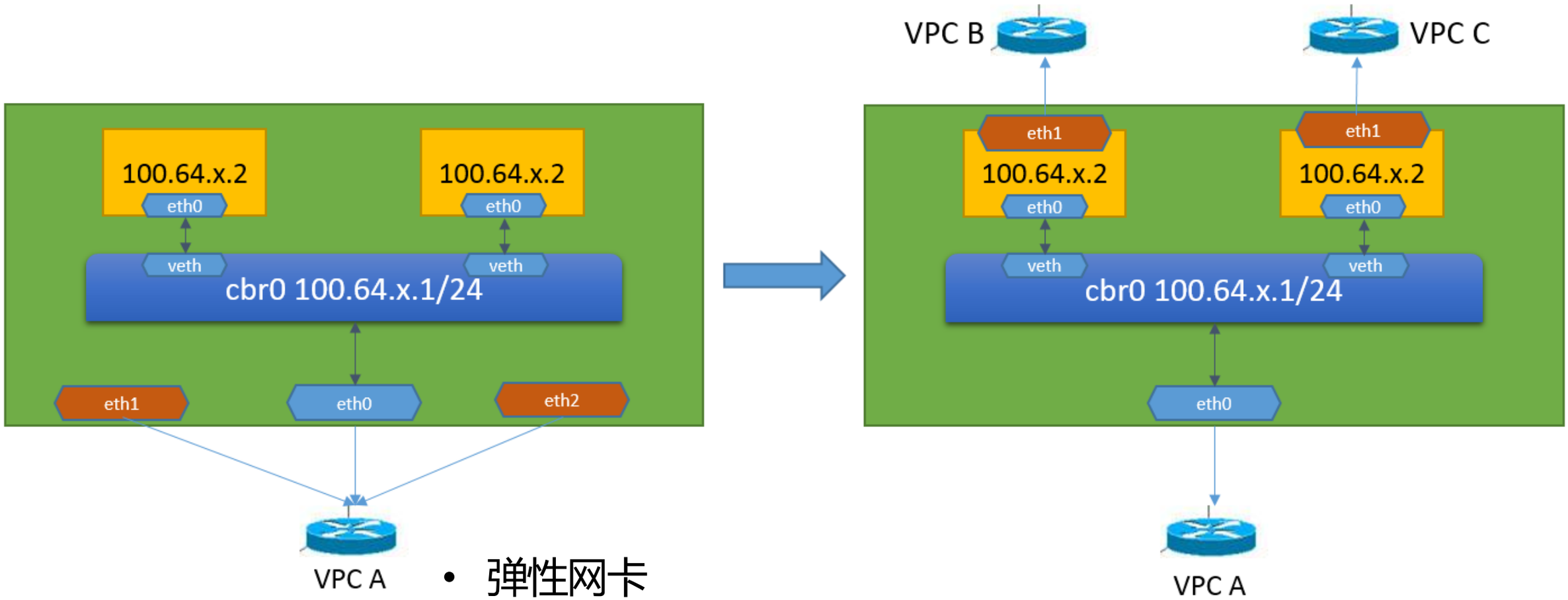
集群监控

新架构

- **单独部署一套k8s集群**
- **用户Master运行为此集群的多个Pod**
- **Pod双网卡**
  - 弹性网卡与用户VPC互通
  - 默认网卡Master Pod间互通



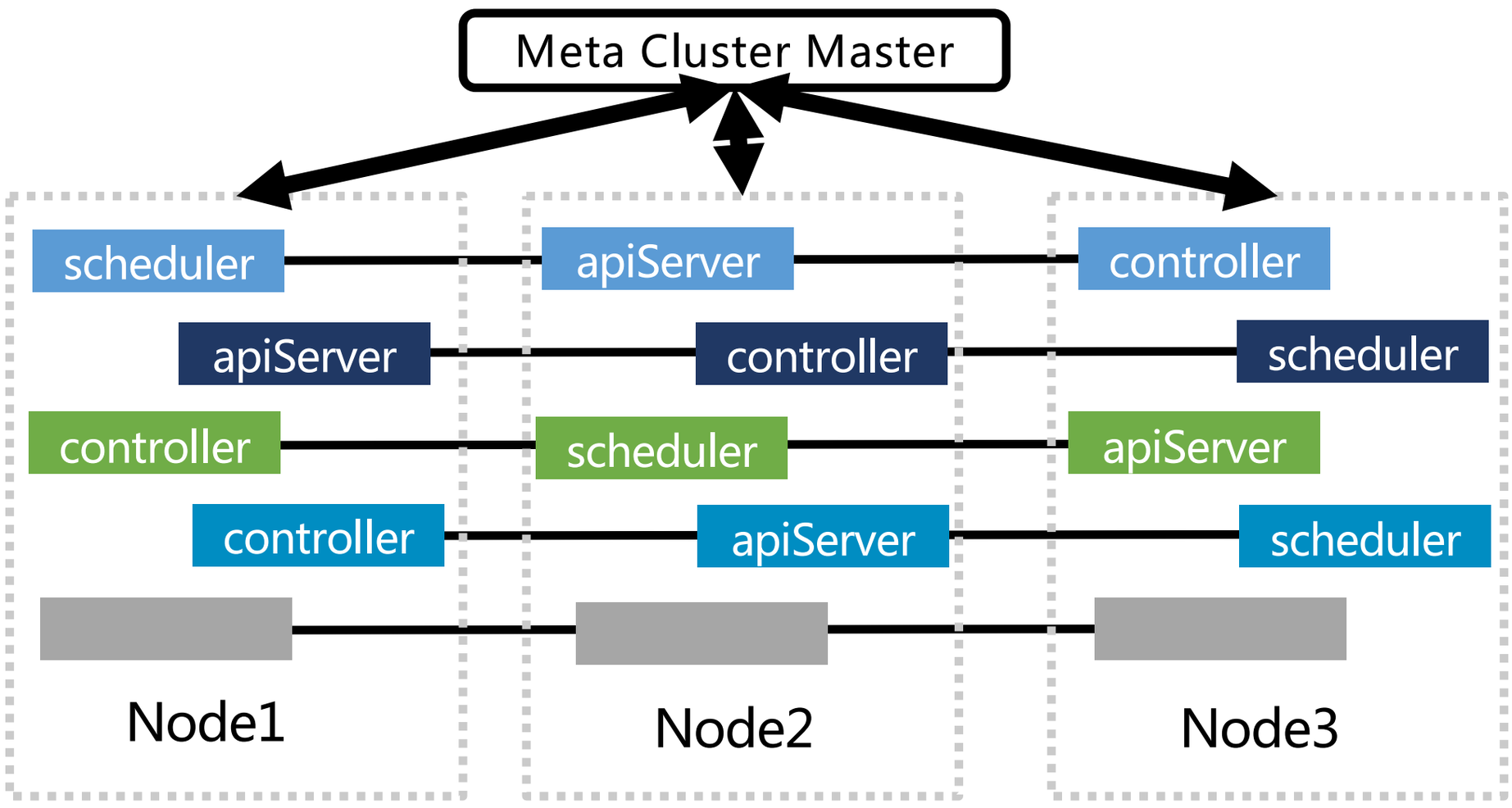
实现



- 弹性网卡
  - 云主机支持绑定不同VPC的弹性网卡
- CNI
  - 弹性网卡移至Pod
  - 对用户VPC的访问走弹性网卡
  - Meta Cluster内部容器互访走eth0(veth)

新架构

- **使用Deployment管理用户集群Master相关组件**
  - HA
  - VPA
  - 成本
  - 一致性



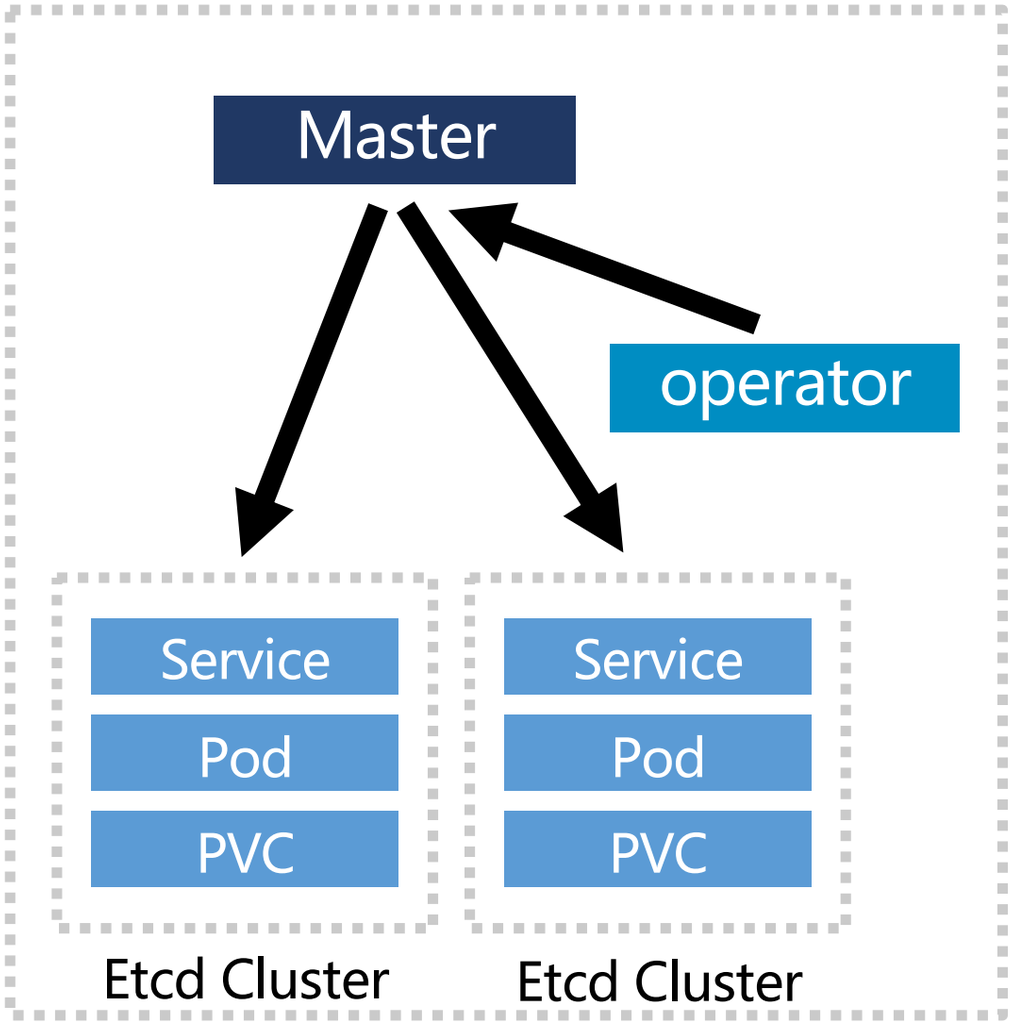
新架构

- **使用应用编排对集群Master相关组件做发布管理**
  - 创建
  - 发布记录
  - 回滚


```yaml
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: kube-controller-manager
  namespace: {{ CLUSTER_NAME }}
spec:
  replicas: 1
  template:
    spec:
      containers:
      - command:
        - controller-manager
        - "--master={{ APISERVER }}"
        - "--allocate-node-cidrs=true"
        - "--cloud-provider=qcloud"
        - "--cluster-cidr={{ CLUSTER_CIDR }}"
        - "--configure-cloud-routes=false"
        - "--leader-elect=true"
        - "--root-ca-file=/etc/kubernetes/secrets/ca.crt"
        - "--service-account-private-key-file=/etc/kubernete
        image: {{ IMAGE }}
        name: kube-controller-manager
        volumeMounts:
        - mountPath: "/etc/kubernetes/secrets"
          name: secrets
        - mountPath: "/etc/ssl/certs"
          name: ssl-host
      volumes:
      - name: secrets
        secret:
          secretName: kube-controller-manager
      - hostPath:
          path: "/usr/share/ca-certificates"
        name: ssl-host
```

新架构

- **使用etcd Operator部署Etcd**
  - 每个集群有独立的Etcd存储
  - SSD Node池
  - 低成本、高可用



新架构

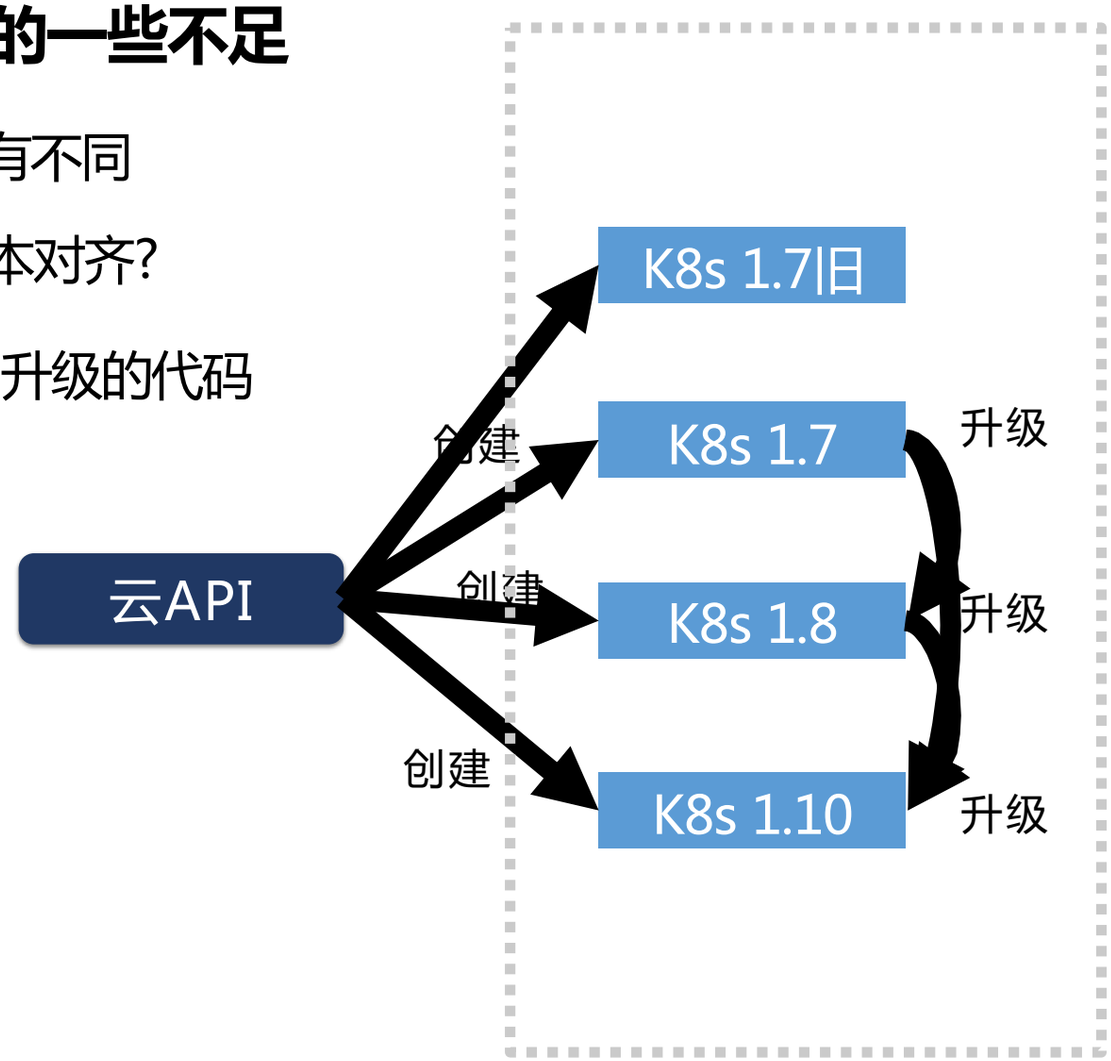
- **监控**
  - **与用户节点一致的容器基础指标监控（CPU,内存，网络）**
    - Pod的VPA（纵向扩容），自动调整Master Pod Request和Limit
  - **事件中心**
    - k8s Event、容器重启、Pod创建超时，etc..

收益

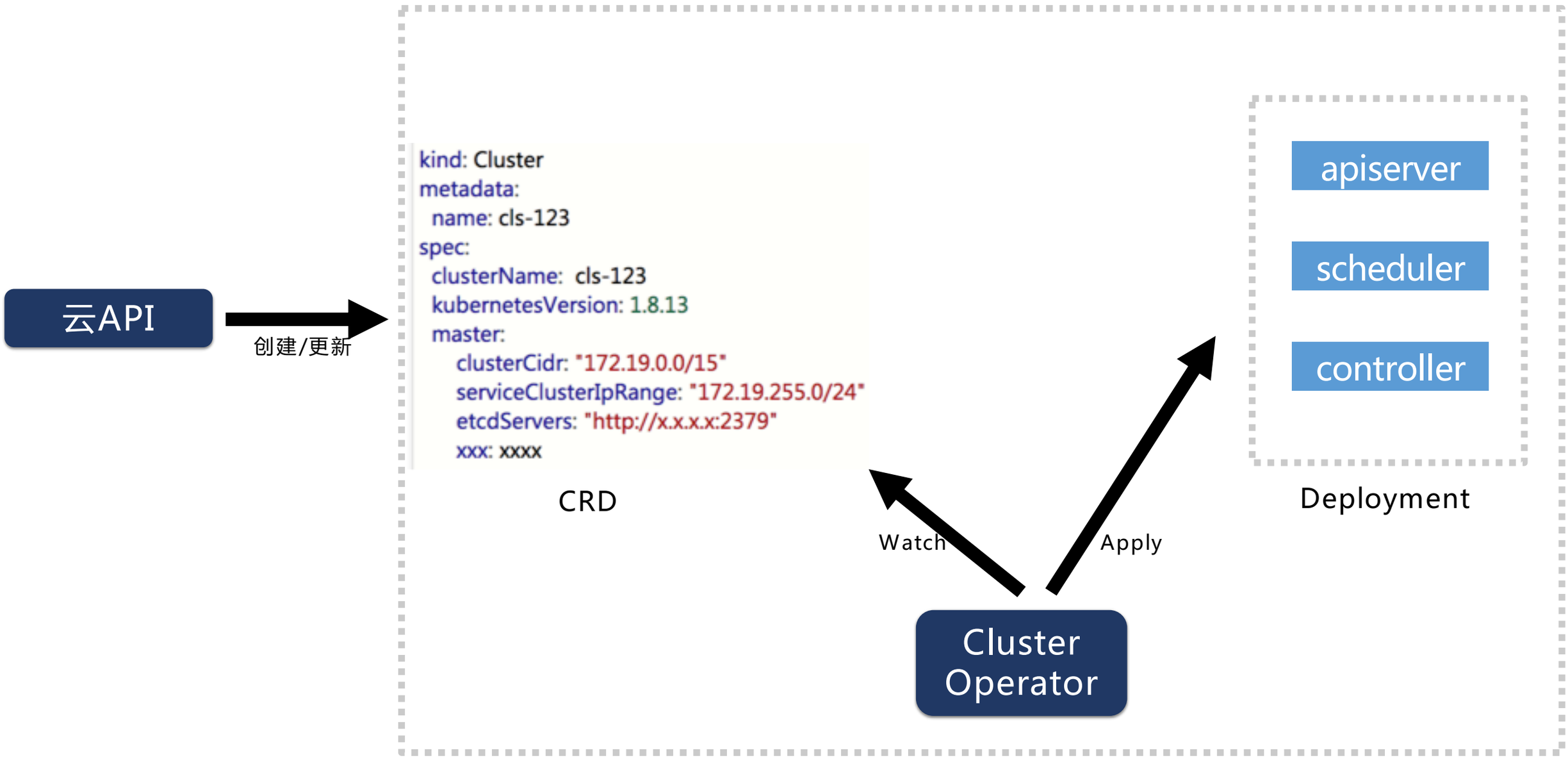
- **Master成本缩减为原来四分之一**
  - 现网有大量冷集群、小集群
- **减少大量无用开发工作**
  - Agent只需初始化、更新Node，无需关注Master相关组件
  - 无需额外实现Master节点HA等
  - 大部分上层功能可作为产品化特性提供给用户（多资源回滚，VPA）
- **统一、一致、简化运维**

Master多版本问题

- **云PI层直接通过Deployment部署Master的一些不足**
  - 版本迭代、手工运维，同一版本k8s参数可能会有不同
  - apiserver , scheduler, controller-manager版本对齐?
  - 除了需要维护创建代码，还需要维护不同版本间升级的代码



k8s Cluster Operator



旧方案架构总览

多Kubernetes运维难点

灵感来源

新方案

总结

总结

- **老架构**
  - Etcd公用
  - 集群Master使用单独VM部署
  - 开发agent来初始化、管理组件
  - 两套监控
- **新架构**
  - 弹性网卡打通网络，实现Master集群化部署
  - 使用统一的部署、监控方式来管理k8s集群
  - 使用Operator管理Master的部署、升级


LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维
LF ASIA, LLC