Alibaba Group 阿里巴巴集团

参数化陷阱与 DSL 缺陷：K8s 声明式应用管理的实践与教训

孙健波

阿里云 技术专家

Geekbang 极客邦科技 | InfoQ 极客传媒

InfoQ官网 全新改版上线

促进软件开发领域知识与创新的传播

关注InfoQ网站
第一时间浏览原创IT新闻资讯

免费下载迷你书
阅读一线开发者的技术干货

Geekbang. 极客邦科技 InfoQ

极客邦科技
会议推荐2019

5月

QCon 北京
全球软件开发大会
大会：5月6–8日
培训：5月9–10日

QCon 广州
全球软件开发大会
培训：5月25–26日
大会：5月27–28日

6月

GTLC GLOBAL TECH LEADERSHIP CONFERENCE 上海
技术领导力峰会
时间：6月14–15日

GMTC 北京
全球大前端技术大会
大会：6月20–21日
培训：6月22–23日

7月

ArchSummit 深圳
全球架构师峰会
大会：7月12–13日
培训：7月14–15日

10月

QCon 上海
全球软件开发大会
大会：10月17–19日
培训：10月20–21日

11月

GMTC 深圳
全球大前端技术大会
大会：11月8–9日
培训：11月10–11日

AiCon 北京
全球人工智能与机器学习大会
大会：11月21–22日
培训：11月23–24日

12月

ArchSummit 北京
全球架构师峰会
大会：12月6–7日
培训：12月8–9日

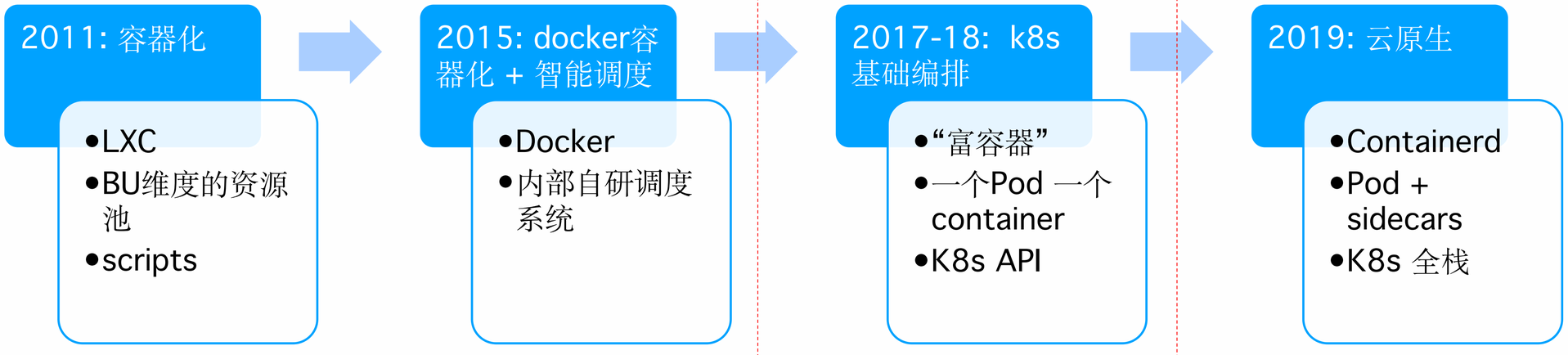
About The SPEAKER

孙健波

- 阿里云 技术专家
- Kubernetes 项目社区成员
- 《Docker容器与容器云》
- PaaS，容器，K8s，日志平台，流式计算。
- jianbo.sjb@alibaba-inc.com

TABLE OF CONTENTS 大 纲

阿里巴巴云原生化路径

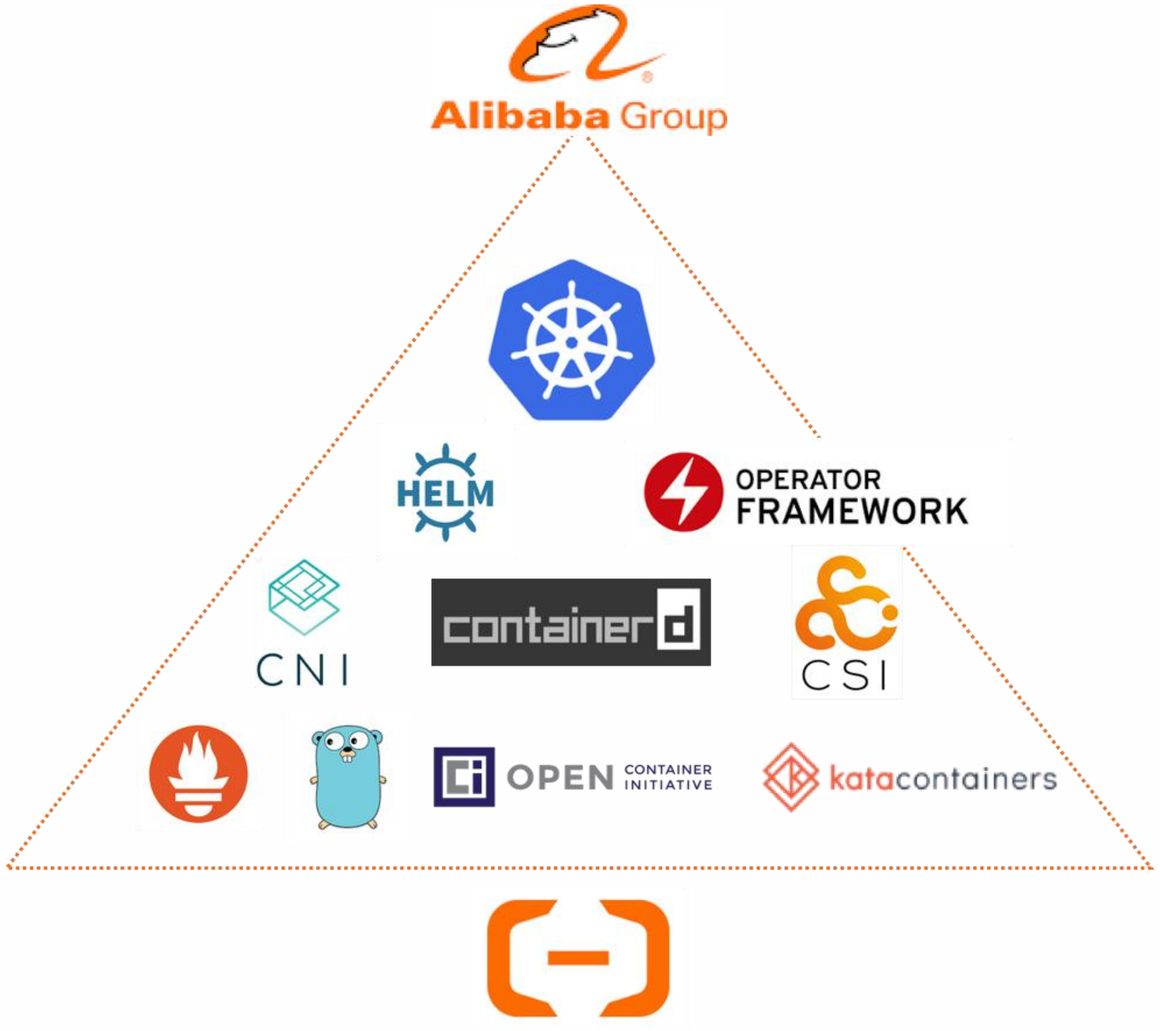


容器化
(围绕容器本身)

探索
(围绕K8s API)

云原生
(标准 + 开放)

当前技术栈



阿里巴巴正在将电商业务全面转型到云上

站在开源界巨人的肩膀上:

- Kubernetes
- Operator Framework
- CNI, CSI, CRI, DevicePlugin …
- Prometheus
- Containerd
  - runC + KataContainers
- DevOps framework from ACK
  - ACK = 阿里巴巴的 Kubernetes 容器服务
- …

整体架构

开发者
/PaaS/DevOps 系统

内部应用管理系统
(监控看板、资源计划、双十一)

搜索引擎、广告、
物流、中间件

Serverless服务、二
方产品、三方系统

声明式 API/GitOps/Kustomize

CRDs + Operators

Kubernetes (阿里服务基础设施)

自定义 Controllers
(Kruise)
kube-controller-manager

Admission Hooks
Aggregator
kube-apiserver

自定义调度器
(Cerebellum)

多租户插件

阿里 ECS 裸金属服务器

containerd

kubelet

阿里 ECS 裸金属服务器

containerd

kubelet

自定义存储接口

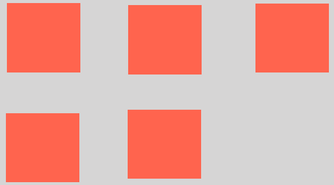
自定义网络接口

云服务商接入

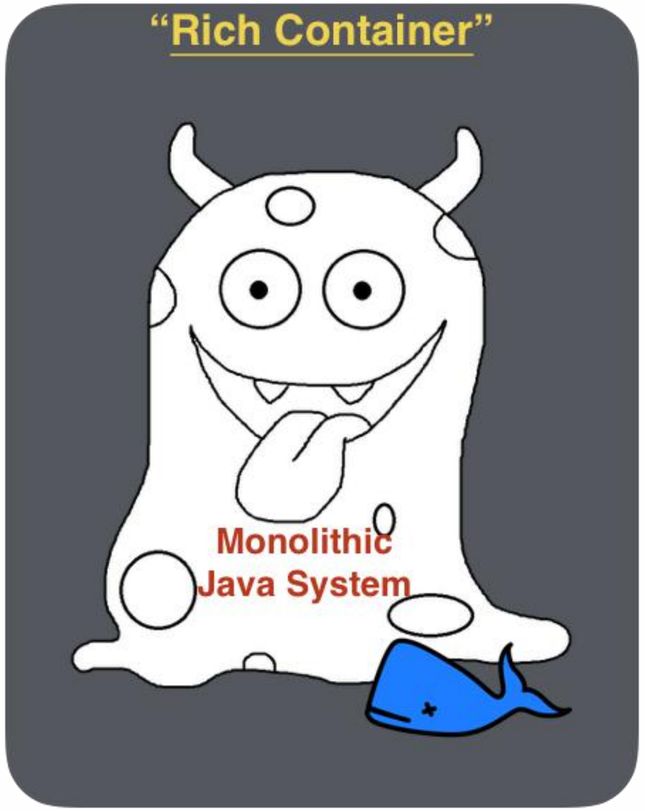
日志

监控

弹性资源池



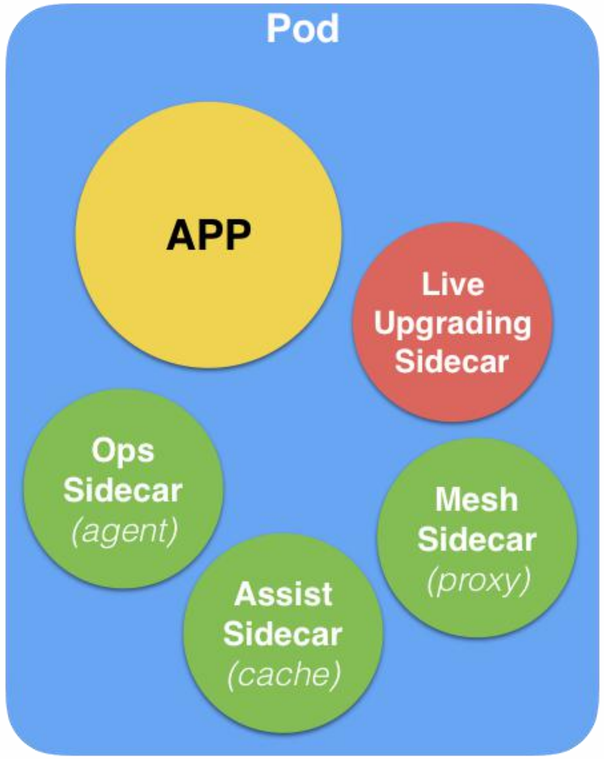
历史遗留问题：富容器



2018 以前

- Java
- 进程(PID 1) 是 Systemd
- 所有组件在同一个容器中 (“富容器”)，组件独立升级
  - app, sshd, log, monitoring, cache, VIP, DNS, proxy, agent, start/stop scripts …
- 传统化的运维工作流
  - 启动容器 -> SSH 进到容器 -> 启动应用
  - 日志、用户数据等在容器内随意存放
- 自研的编排、调度系统

从“容器”到“应用”



- 共享数据卷
- 不同资源不同QoS
- 可控的生命周期管理和健康检查

```
apiVersion: v1
kind: Pod
spec:
  containers:
  - env:
    - name: ali_start_app
      value: “no”
    name: main
    lifecycle:
      postStart:
        exec:
          command:
          - /bin/sh
          - -c
          - for i in $(seq 1 60); do [ -x /home/admin/.start ] && break ; sleep 5
            ; done; sudo -u admin /home/admin/.start>/var/log/kubeapp/start.log 2>&1
            && sudo -u admin /home/admin/health.sh>>/var/log/kubeapp/start.log 2>&1
      preStop:
        exec:
          command:
          - /bin/sh
          - -c
          - sudo -u admin /home/admin/stop.sh>/var/log/kubeapp/stop.log 2>&1
    livenessProbe:
      exec:
          command:
          - /bin/sh
          - -c
          - sudo -u admin /home/admin/health.sh>/var/log/kubeapp/health.log 2>&1
        initialDelaySeconds: 20
        periodSeconds: 60
        timeoutSeconds: 20
```

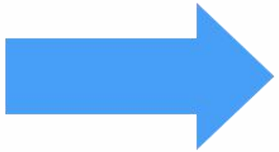
App start script

App stop script

Health check

云原生化关键节点：轻量级容器化

- Kubernetes 语境中的“应用” = **“应用配置”**
  - **该软件制品的存放路径?**
    - Container Image Registry
  - **该软件如何运行、发布和更新?**
    - Workloads (StatefulSet, Deployment …)
  - **如何访问该软件?**
    - Services, Ingresses
  - **访问该软件的网络约束?**
    - NetworkPolicies



应用描述文件

- Helm 应用（Charts）
- 原生 K8s YAML
- App Definition

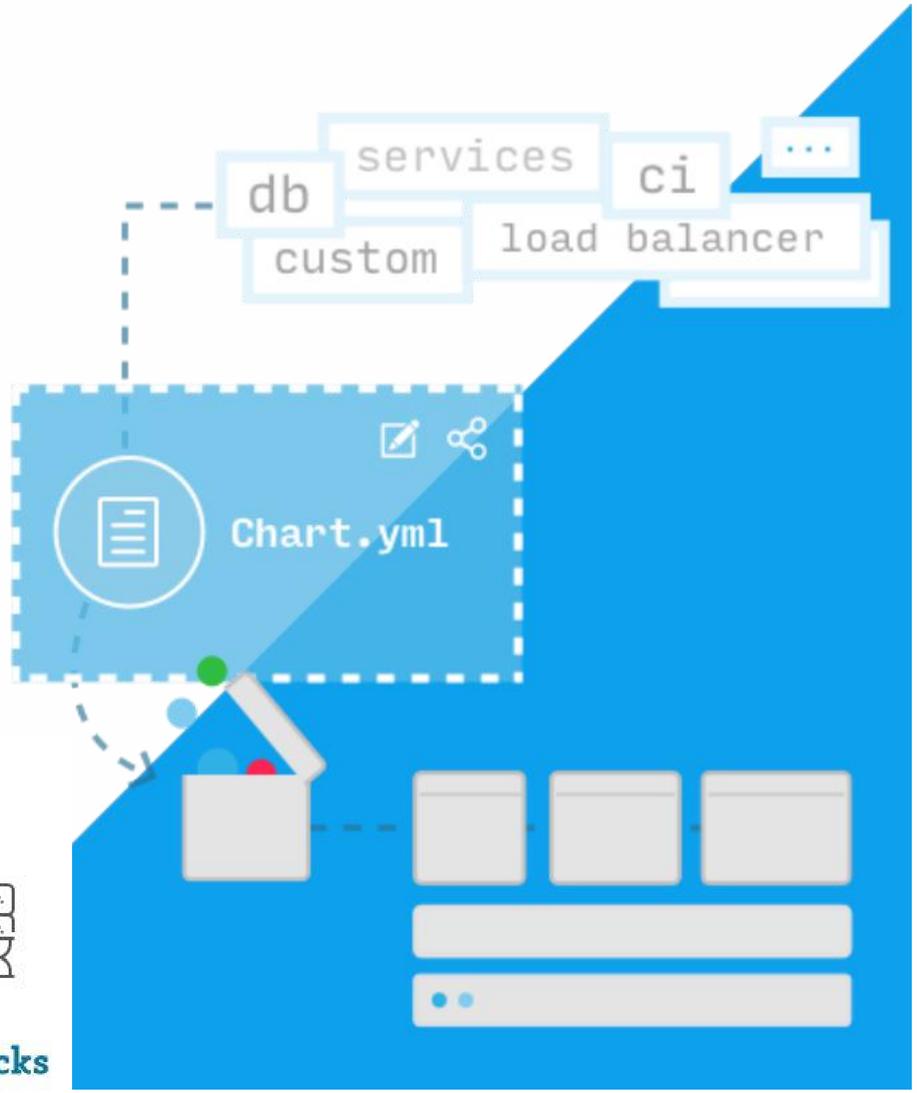
```
$ helm install _APP_
$ kubectl apply -f _YAML_
```

标准应用定义

- **Helm ?**

我们在使用 Charts

解决了大部分应用定义的问题。



参数化陷阱 与 DSL 缺陷

- Helm 的好处

1. 应用管理功能完善，尤其是依赖管理和回滚
2. 端到端的使用体验，快速达成部署的目的。
3. 包管理器：解决了应用打包问题

- Helm 的问题

1. Templating 与 Kubernetes 的声明式 API （Patch 方式）不一致
2. 服务端 Tiller （Helm v3会解决）
3. “二层 API”： Helm v3 引入 LUA（或任何类型的 DSL）
4. “一个不能直接使用的模板文件”+“一个没有实际意义的 values 文件”

Kustomize 是什么?

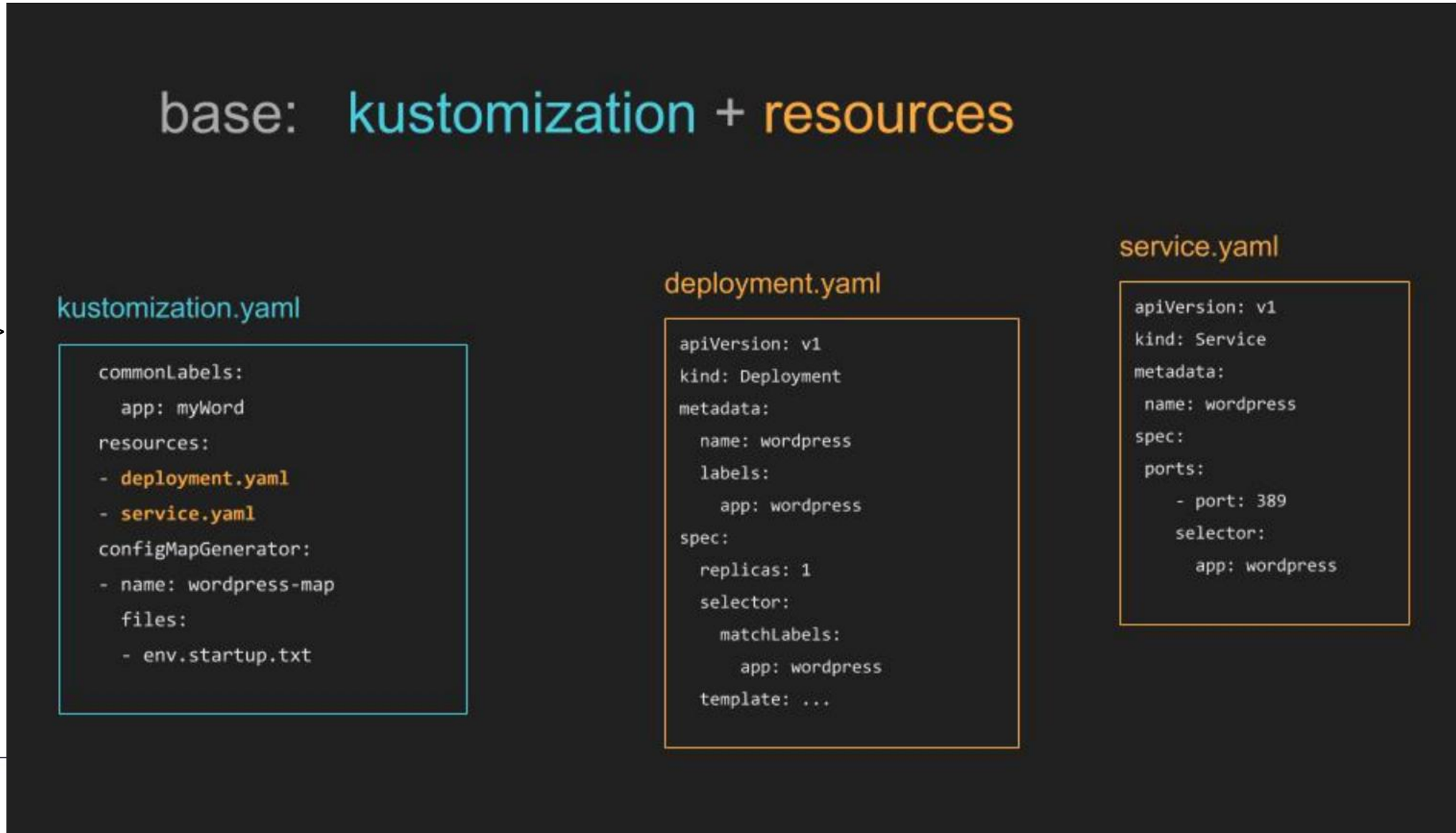
V1.14

查看输出

$ kubectl kustomize <dir>

执行

$ kubectl –k <dir>



Kustomize 强大的overlay

overlay: kustomization + patches + more resources

(referencing a base)

kustomization.yaml

```
namePrefix: prod-
commonLabels:
  variant: prod
commonAnnotations:
  note: Hello, I am production!
bases:
- ../../base
patches:
- replica_count.yaml
- cpu_count.yaml
```

replica_count.yaml

```
apiVersion: v1
kind: Deployment
metadata:
  name: wordpress
spec:
  replicas: 80
```

cpu_count.yaml

```
apiVersion: v1
kind: Deployment
metadata:
  name: wordpress
spec:
  template:
    spec:
      containers:
      - name: my-container
        resources:
          limits:
            cpu:  7000m
```

Helm + Kustomize

Kustomize 的好处

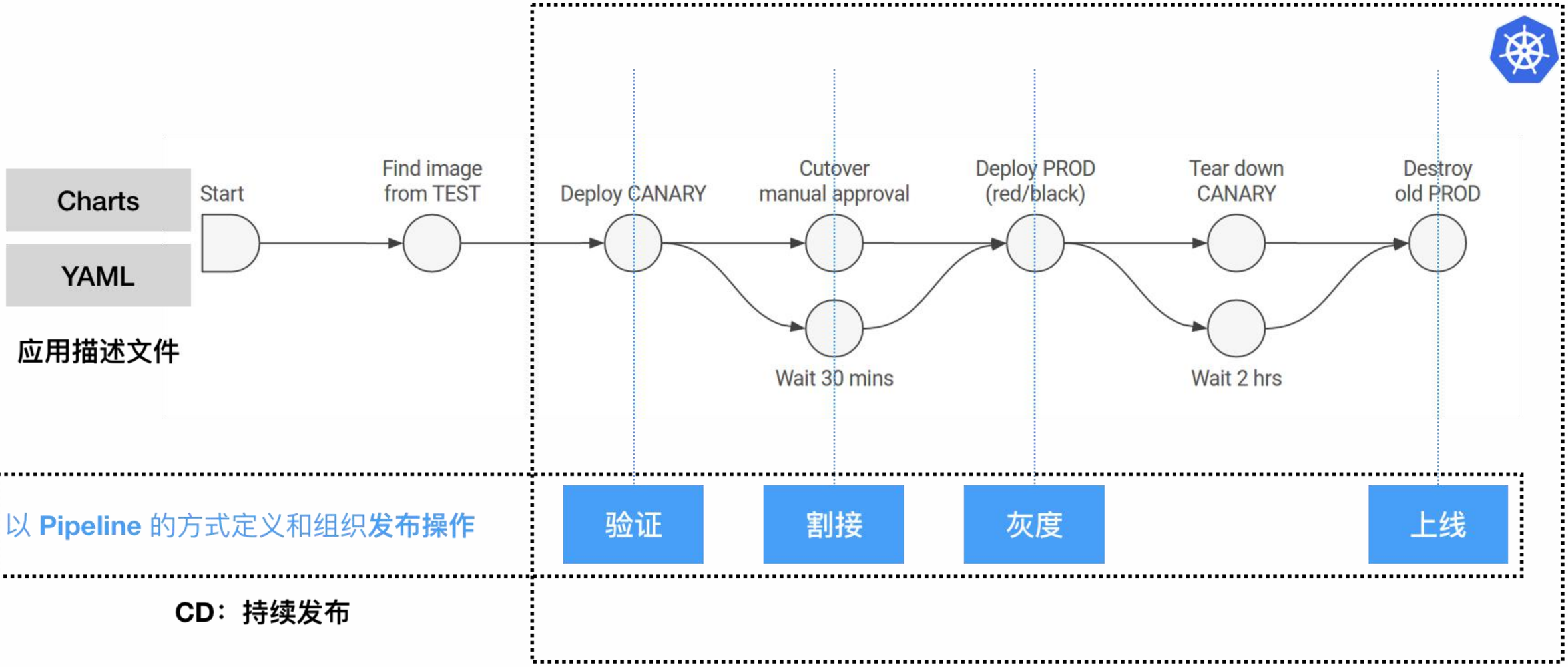
1. 定制 YAML 的体验很好。
2. 只关注 Templating 环节，从设计上就是被集成的位置而非盖在 Kubernetes 之上“二层 API”。
3. 与声明式 API 的体感匹配，没有割裂感。

Kustomize 的问题

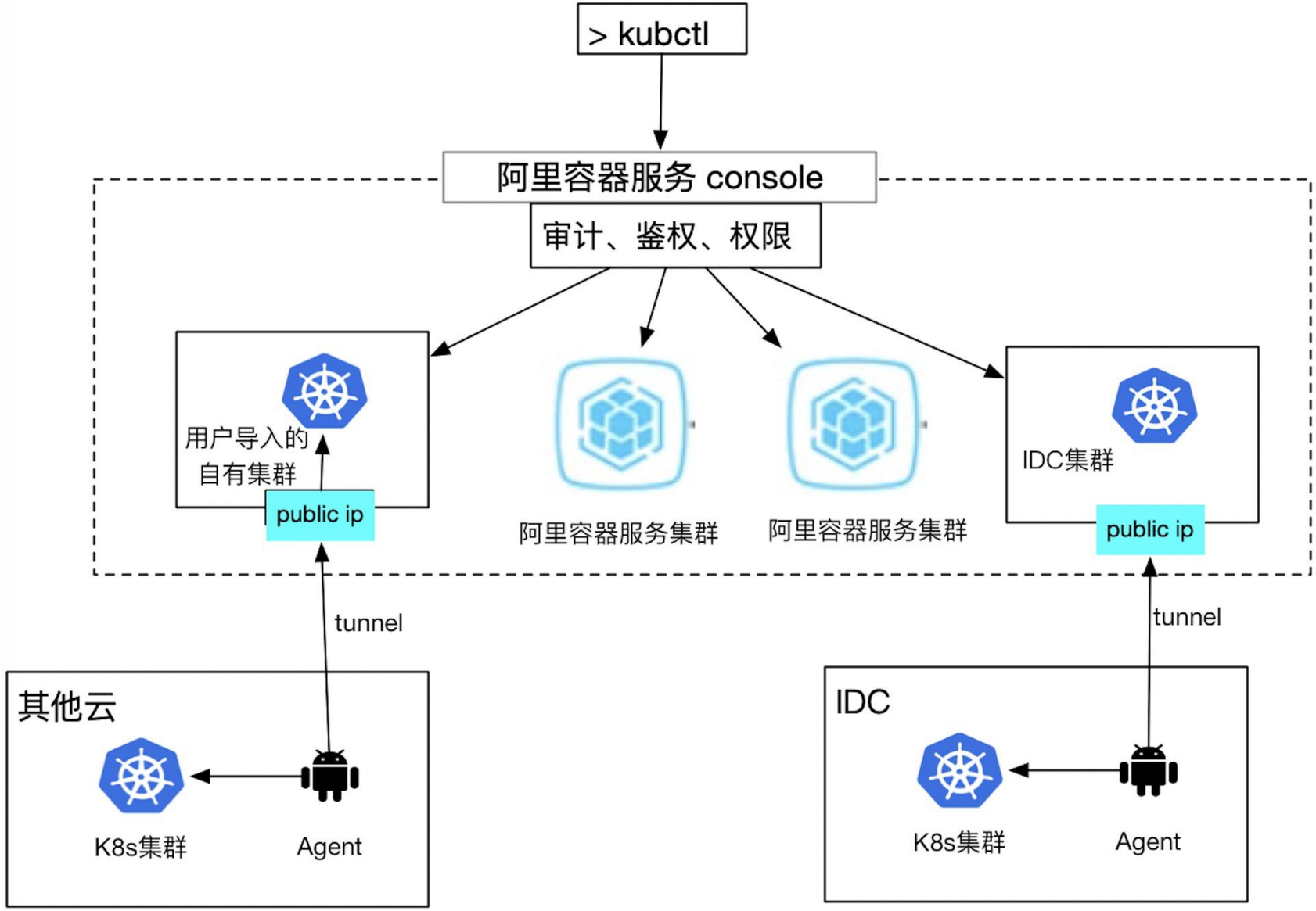
1. 管理 Kustomize 文件不比管理 Charts 文件轻松。
2. 不提供其他能力，需要与更多的项目配合使用甚至需要用户自己 DIY pipeline。

Kustomize 解决了我们在规模化场景下修改 YAML 文件的难题

从应用到应用交付


Charts
YAML
应用描述文件
Start
Find image from TEST
Deploy CANARY
Cutover manual approval
Deploy PROD (red/black)
Tear down CANARY
Destroy old PROD
Wait 30 mins
Wait 2 hrs
以 Pipeline 的方式定义和组织发布操作
验证
割接
灰度
上线
CD：持续发布

混合云、多云

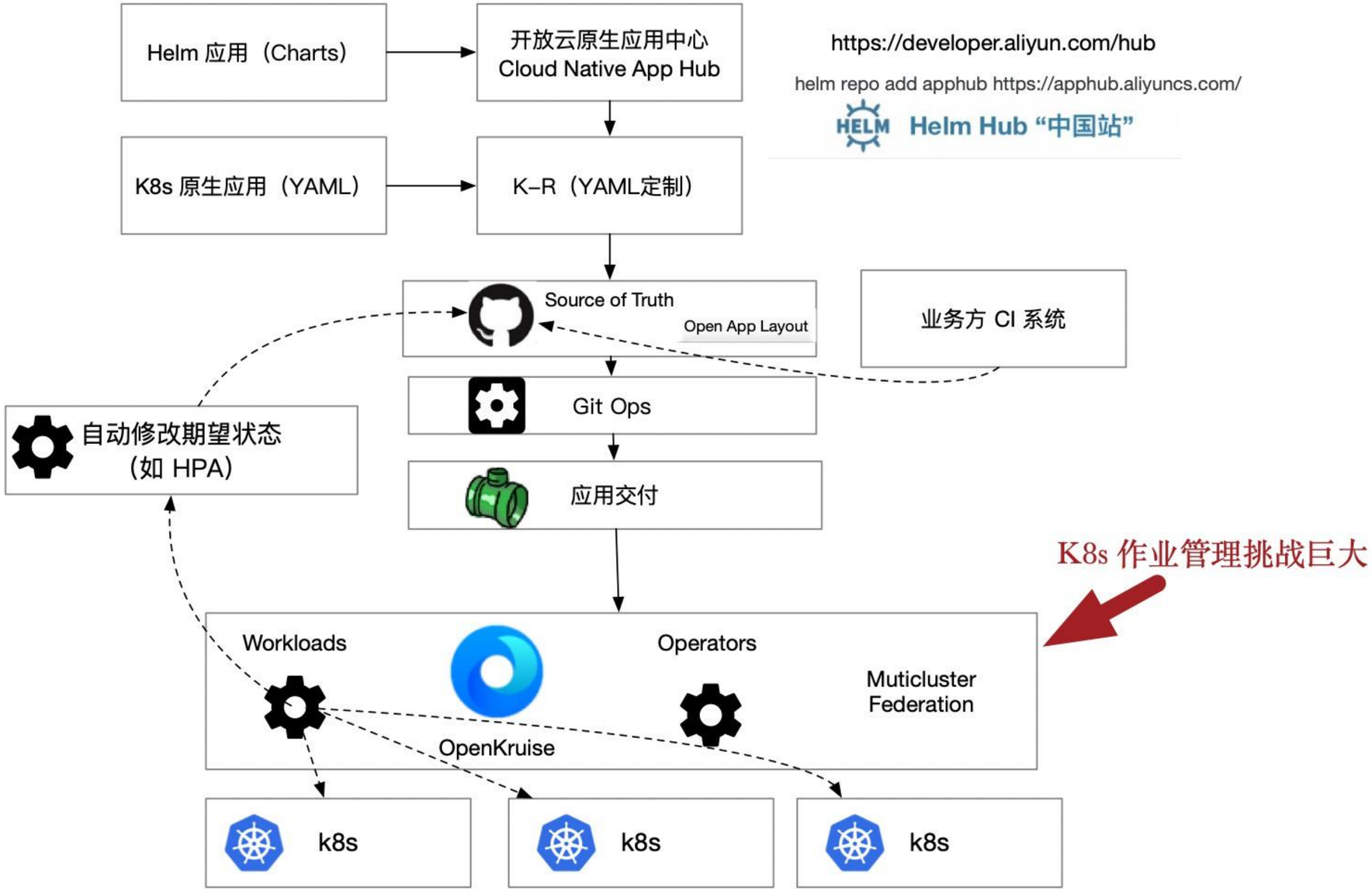

> kubctl
阿里容器服务 console
审计、鉴权、权限
用户导入的
自有集群
public ip
阿里容器服务集群
阿里容器服务集群
IDC集群
public ip
tunnel
tunnel
其他云
K8s集群
Agent
IDC
K8s集群
Agent

应用是一组带有运维属性的、且互相之间具有关联关系的组件集合。

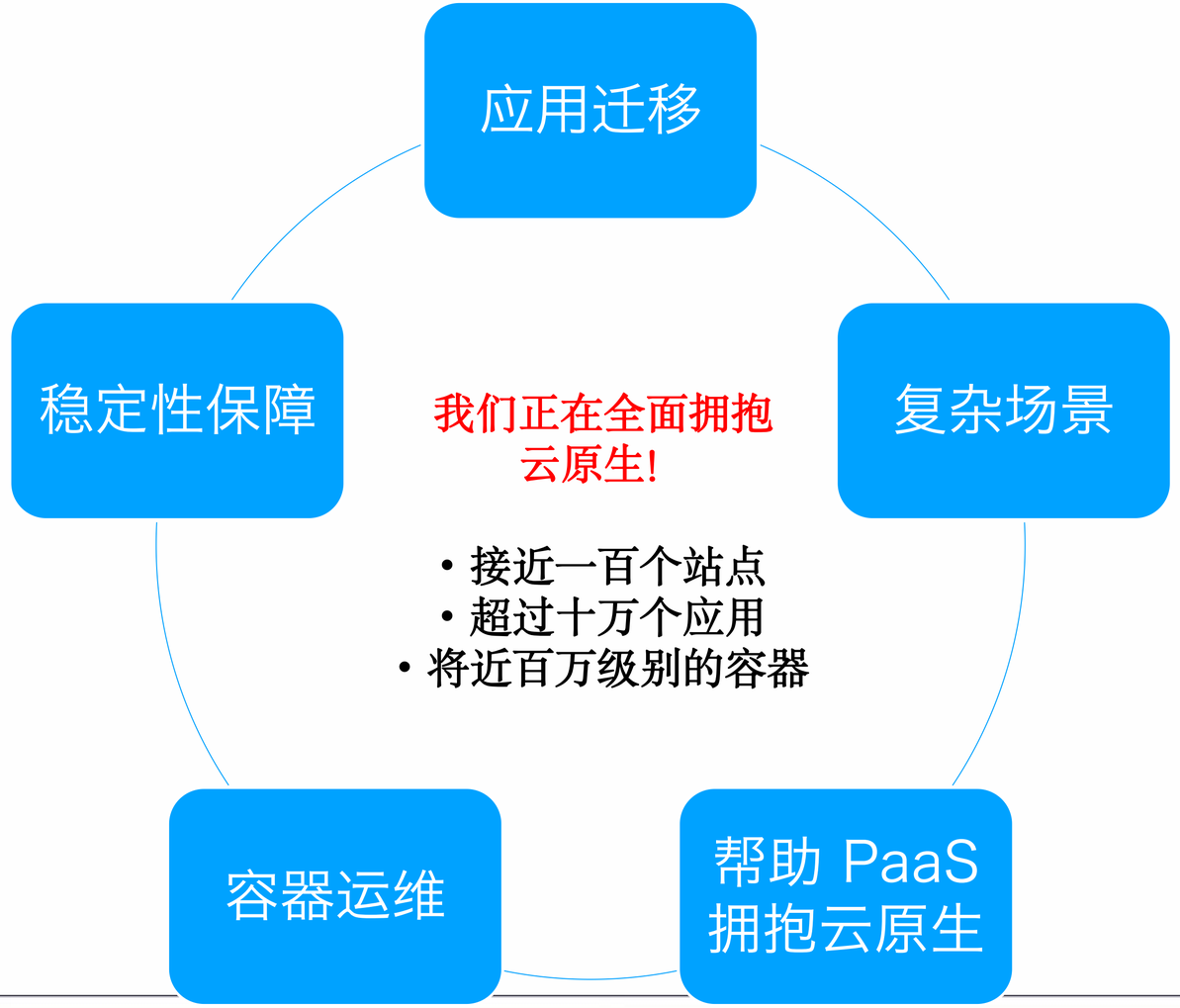
1. 角色：开发者、应用运维、基础设施运维
2. 组件：可运行单元
3. 工作负载：组件能运行的不同工作负载。
4. 属性：应用参数，运维关心的参数。
5. 应用边界：组件间共有的属性、组件间的依赖关系。
6. 操作配置：组件实例、他们的属性、作用域、参数。

标准应用定义，即将开源!

云原生应用架构


Helm 应用（Charts）
开放云原生应用中心
Cloud Native App Hub
https://developer.aliyun.com/hub
helm repo add apphub https://apphub.aliyuncs.com/
Helm Hub “中国站”
K8s 原生应用（YAML）
K–R（YAML定制）
Source of Truth
Open App Layout
业务方 CI 系统
Git Ops
自动修改期望状态
（如 HPA）
应用交付
K8s 作业管理挑战巨大
Workloads
Operators
Muticluster
Federation
OpenKruise
k8s
k8s
k8s

阿里巴巴 K8s 作业管理的多重挑战



举例：“双十一”大促

在阿里双十一大促前，我们要准备：

1. 为上千个应用创建 10万+ Pods
2. 通过一个离线调度器，计算node数量，提前创建pod
3. 通过上述方式，提升资源利用率、节省CPU、编排隔离/混部应用 等等

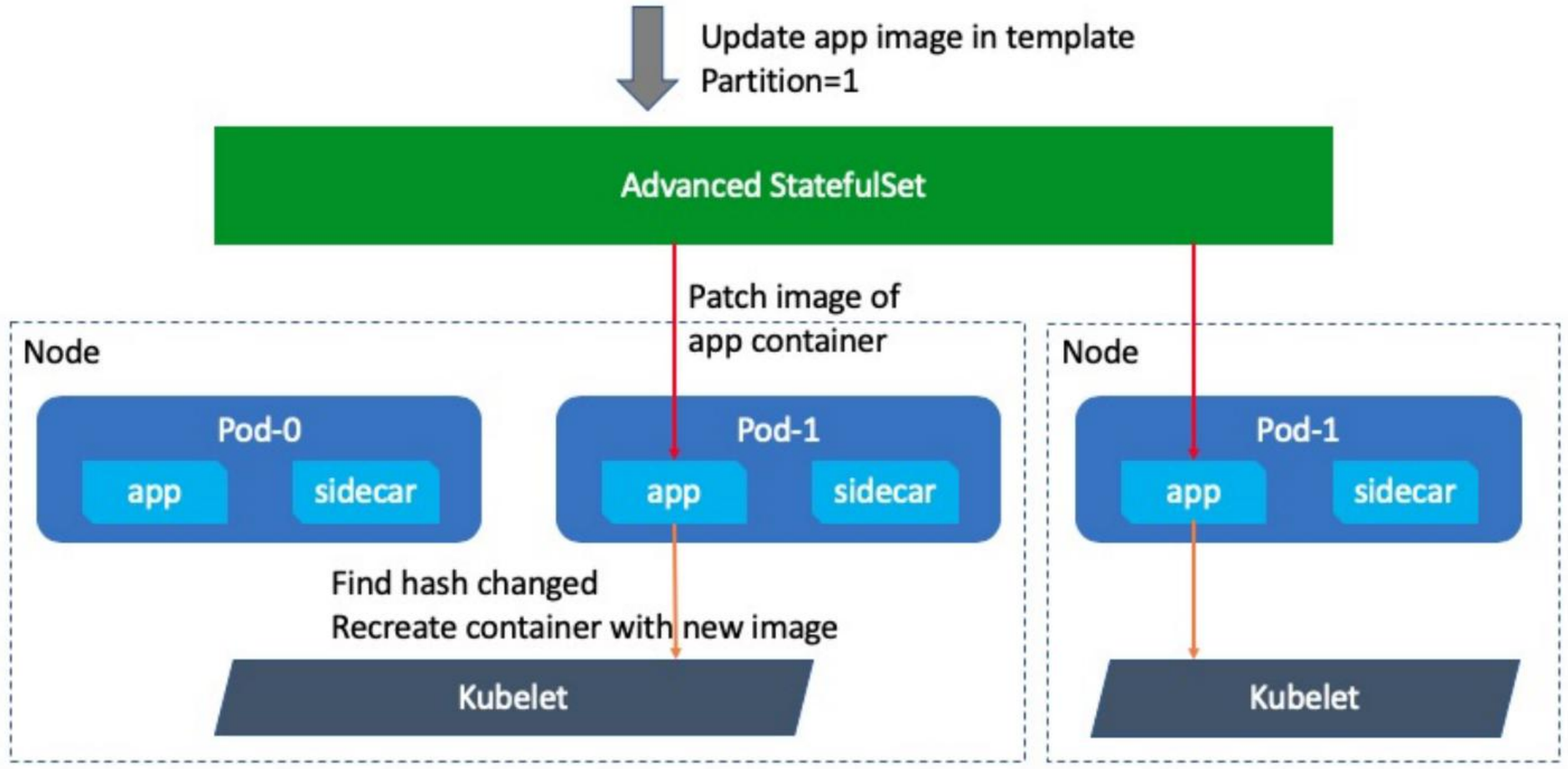
在滚动升级时，默认情况下所有的工作负载都会重新创建Pods，可用策略：

- maxSurge/maxUnavailable/partition/…

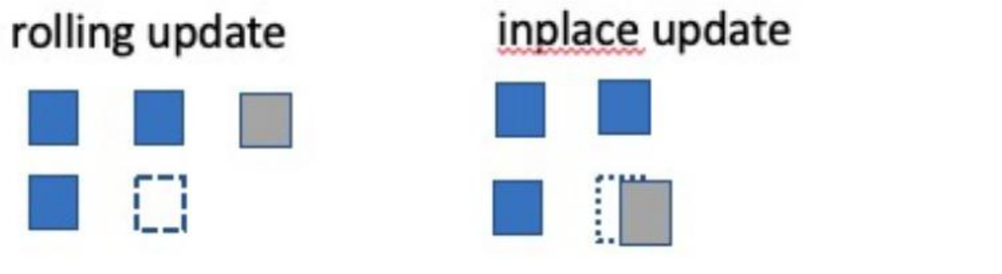
问题在哪？

- 如果我们用默认的更新策略，预期的功能完全不成立：
- 拓扑逻辑变化、镜像预热没了、额外开销增加、资源分配混乱 …

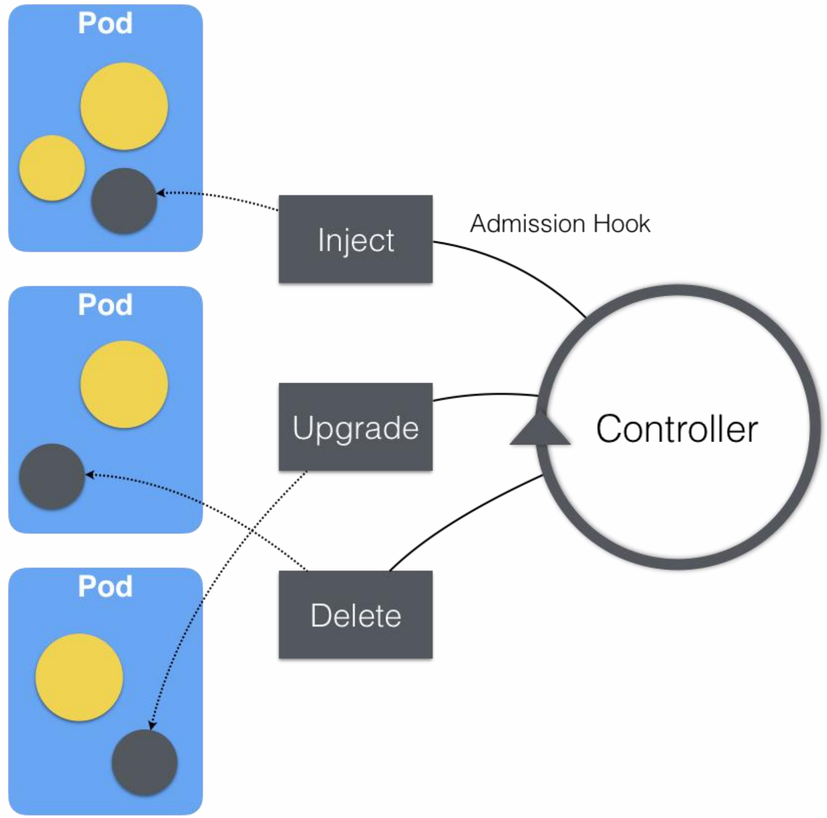
更“高阶”的 StatefulSet需要什么能力?



发布策略：原地升级、SidecarSet、BroadcastJob



	Recreate update	InPlace update
集群确定性		👍
镜像下载性能		👍
资源本身要求		👍
服务重新调度和注册发现		👍
自动恢复	👍	
全量字段更新	👍	



一起进入云原生架构时代!

- 阿里云与 CNCF 联合制作了《CNCF x Alibaba 云原生技术公开课》，从 K8s 基础开始到进阶再到实践，循序渐进为你一一分解，欢迎即刻开始学习。

- OpenKruise 开源地址
  https://github.com/openkruise/kruise

- Cloud Native App Hub
  https://developer.aliyun.com/hub

极客时间全部课程任学

喊老板来买单！

- 精选 13+ 热门职位的学习路径，包括架构、运维、前端工程师等
- 根据不同技术岗位能力模型匹配合适的课程
- 一键设置购买条件，成员按需选课，自主制定学习计划
- 享充值满赠优惠，帮老板省钱，团队免费学习

立即申请

Geekbang 极客邦科技 | TGO 鲲鹏會

TGO鲲鹏会

汇聚全球科技领导者的高端社群

全球12大城市

850+高端科技领导者

使命 Mission

为社会输送更多优秀的科技领导者

愿景 Vision

构建全球领先的有技术背景优秀人才的学习成长平台

扫描二维码，了解更多内容

Alibaba Group
阿里巴巴集团

THANKS

Geekbang. 极客邦科技 InfoQ