阿里巴巴云原生专场

# 阿里巴巴云原生容器基础设施运维实践

KubeNode：manage large scale Kubernetes nodes in Cloud-Native fashion

演讲人：周 涛 (广 侯)

# 关于我

- 2017年加入阿里，负责阿里巴巴数十万集群节点运维管控系统研发，参与历年双11大促。
- 随着集团上云，目前在阿里云云原生应用平台， 负责Serverless节点管理平台研发。

🔥 We are hiring🔥 (WeChat/TG: espacewalker)

钉钉答疑群：

# Agenda

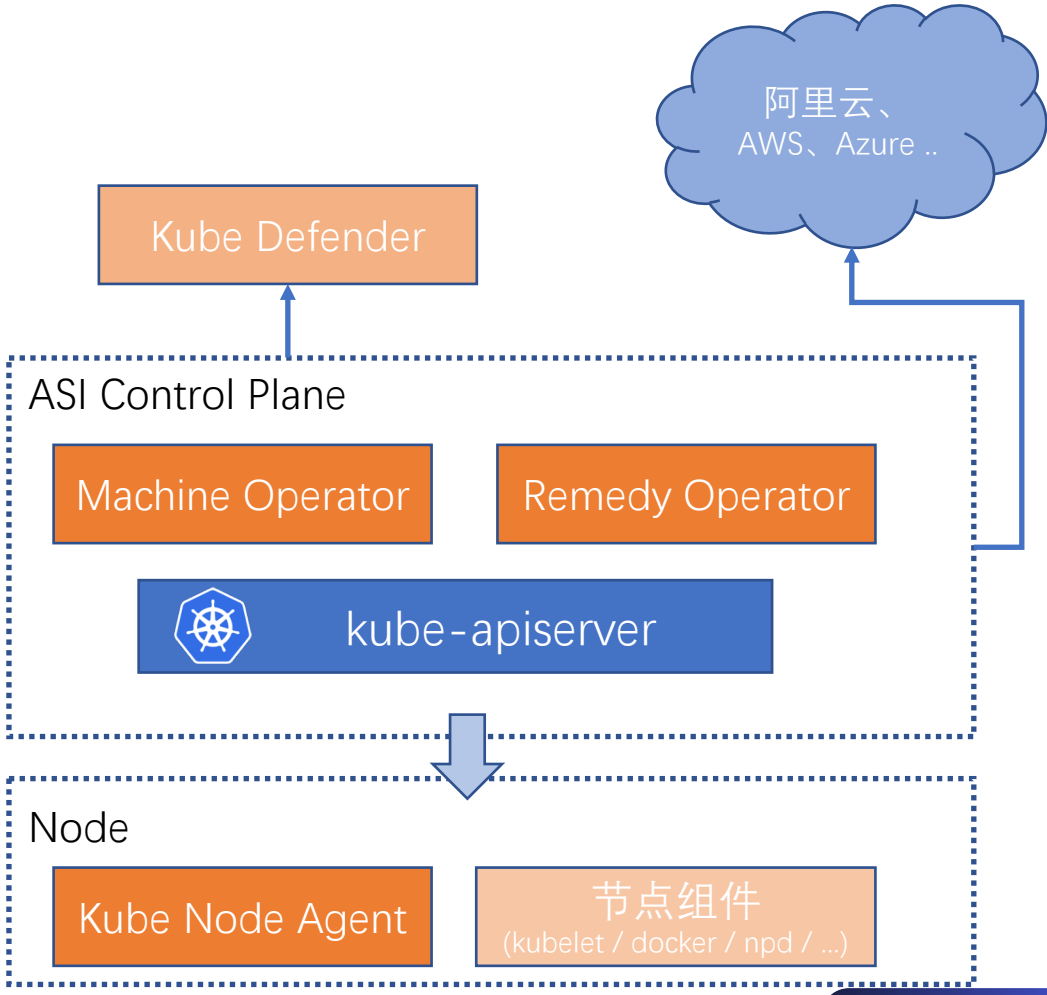
- 阿里巴巴节点运维的挑战
- KubeNode：云原生节点运维底座
- 未来展望

# 阿里巴巴节点运维的挑战

- 规模大
  - 数百ASI集群 (Ali Serverless Infra, ACK + Ali addon)
  - 数十万节点 (单集群节点最多~10k台)
  - 数万应用
  - 数百万容器
- 环境复杂
  - x86 / ARM / GPU / FPGA
  - 在线 (应用类型差异大)、混部、安全容器
- 稳定性要求高
  - 在线业务延迟、抖动敏感
  - 宕机、夯机业务无感知

# KubeNode：阿里巴巴云原生节点底座

- What & Why
  - 以云原生方式管理节点生命周期及节点组件
  - 申明式、面向终态
- 组成：
  - 中心端：
    - Machine Operator： 节点及组件管理
    - Remedy Operator： 节点故障自愈
  - 节点侧：
    - Kube Node Agent：单机 agent
  - 配套组件：
    - Kube Defender 统一风控
    - NPD: 单节点故障检测



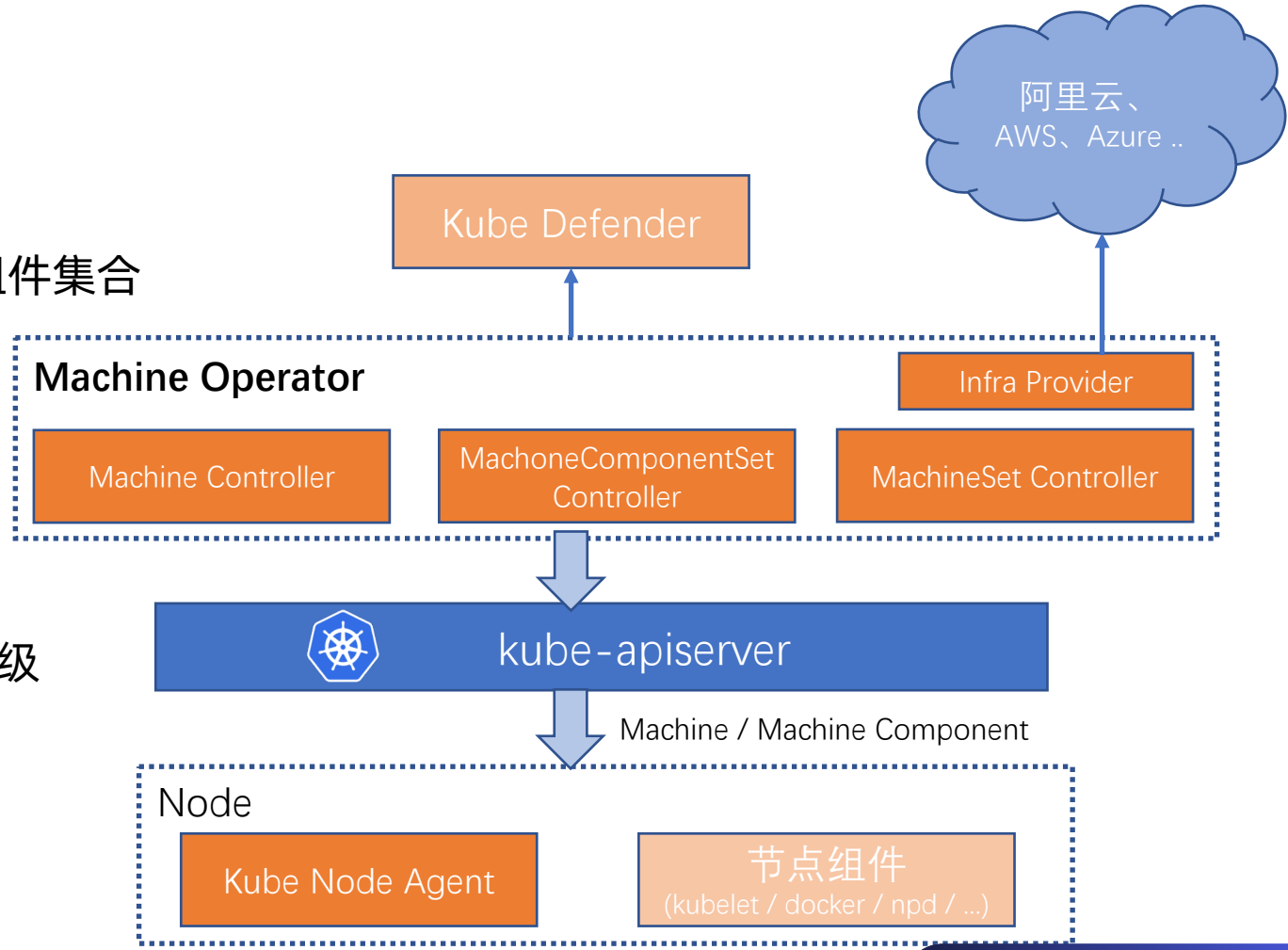
# KubeNode 和社区项目关系

- github.com/kube-node
  - 不相关，该项目2018年初已停止
- ClusterAPI
  - KubeNode 可以作为 ClusterAPI 节点终态的补充
  - 功能对比：

| | Cluster API | KubeNode |
|---|---|---|
| 集群 Provision | Yes | No |
| 节点 Provision | Yes | Yes |
| 节点组件终态 | No | Yes |
| 节点故障自愈 | Yes (simple) | Yes (full, rule based) |

# KubeNode - Machine Operator

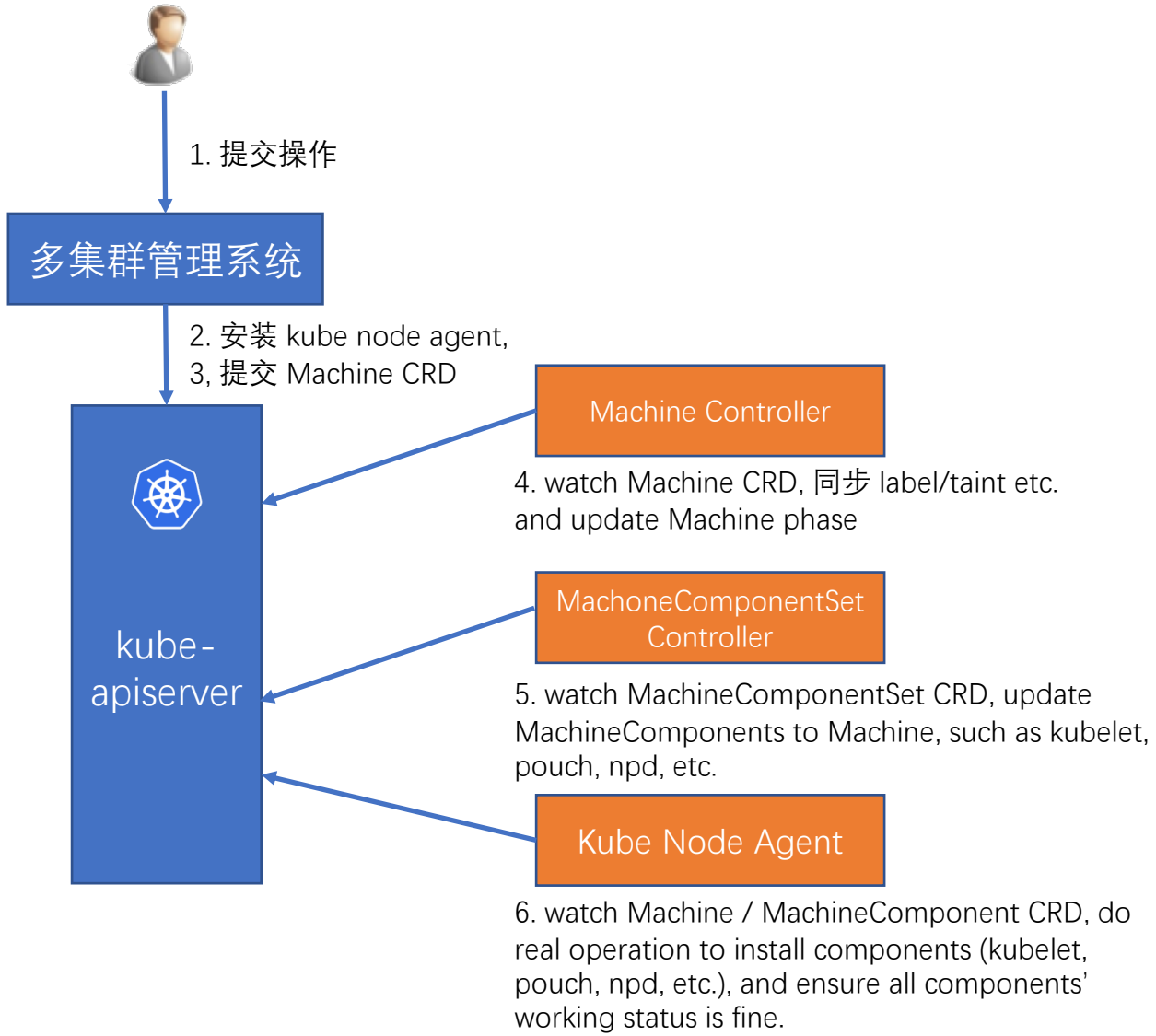
- CRDs
  - Machine: 节点元信息
  - MachineSet (MS): 节点集合
  - MachineComponentSet (MCS): 节点组件集合
  - MachineComponent (MC): 节点组件
- Controllers
  - MS controller: 节点 provision
  - MCS controller: 节点组件分批安装、升级
  - Infra Provider: 对接云厂商 OpenAPI
- Kube Node Agent
  - 单机组件安装、升级、终态维持



# Use Case: 节点导入

- k8s 扩展 CRD 描述节点及组件
  - Machine
  - MachineComponent
  - MachineComponentSet

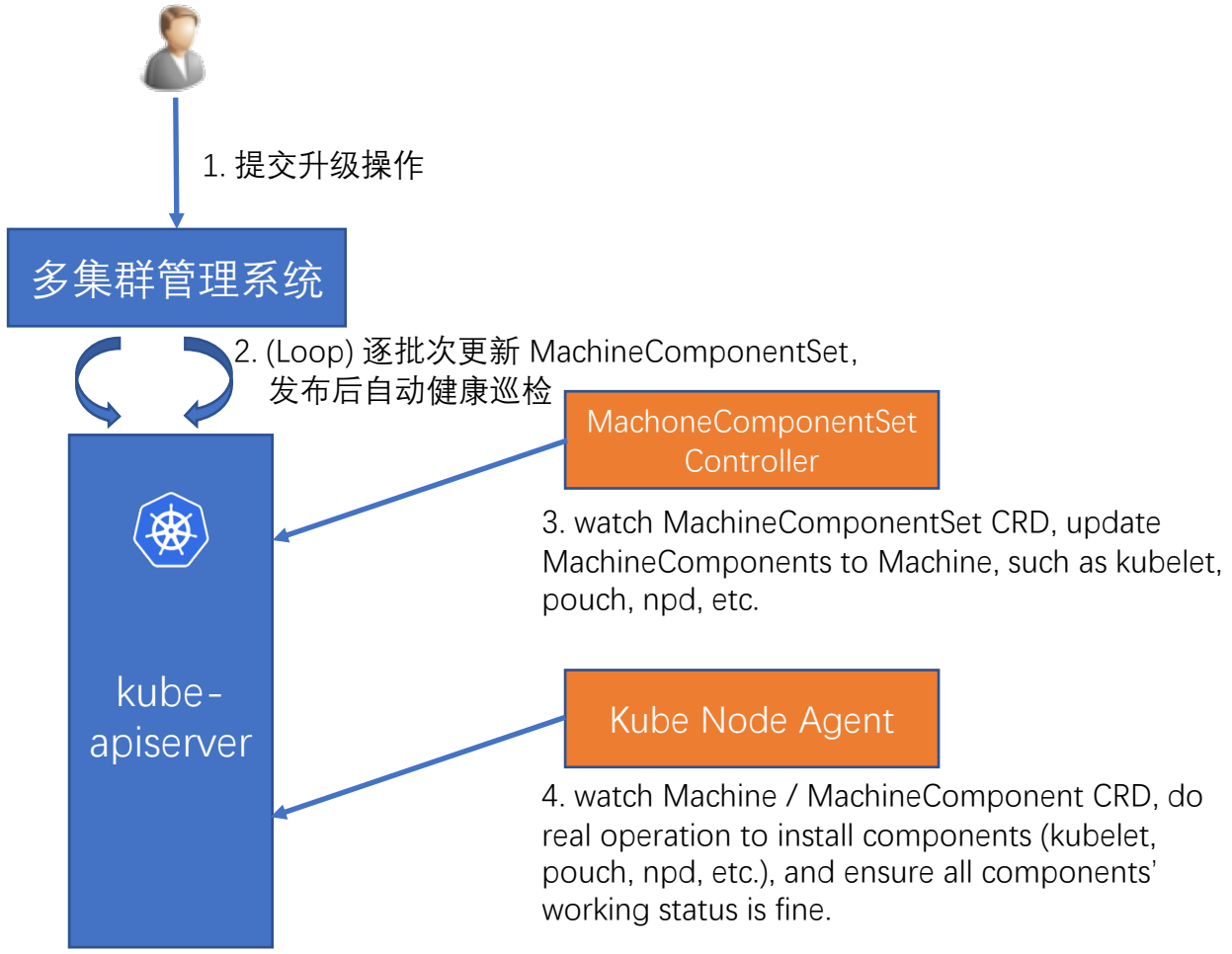
- 节点组件确保终态一致
  - version
  - config
  - status



# Use Case: 组件升级

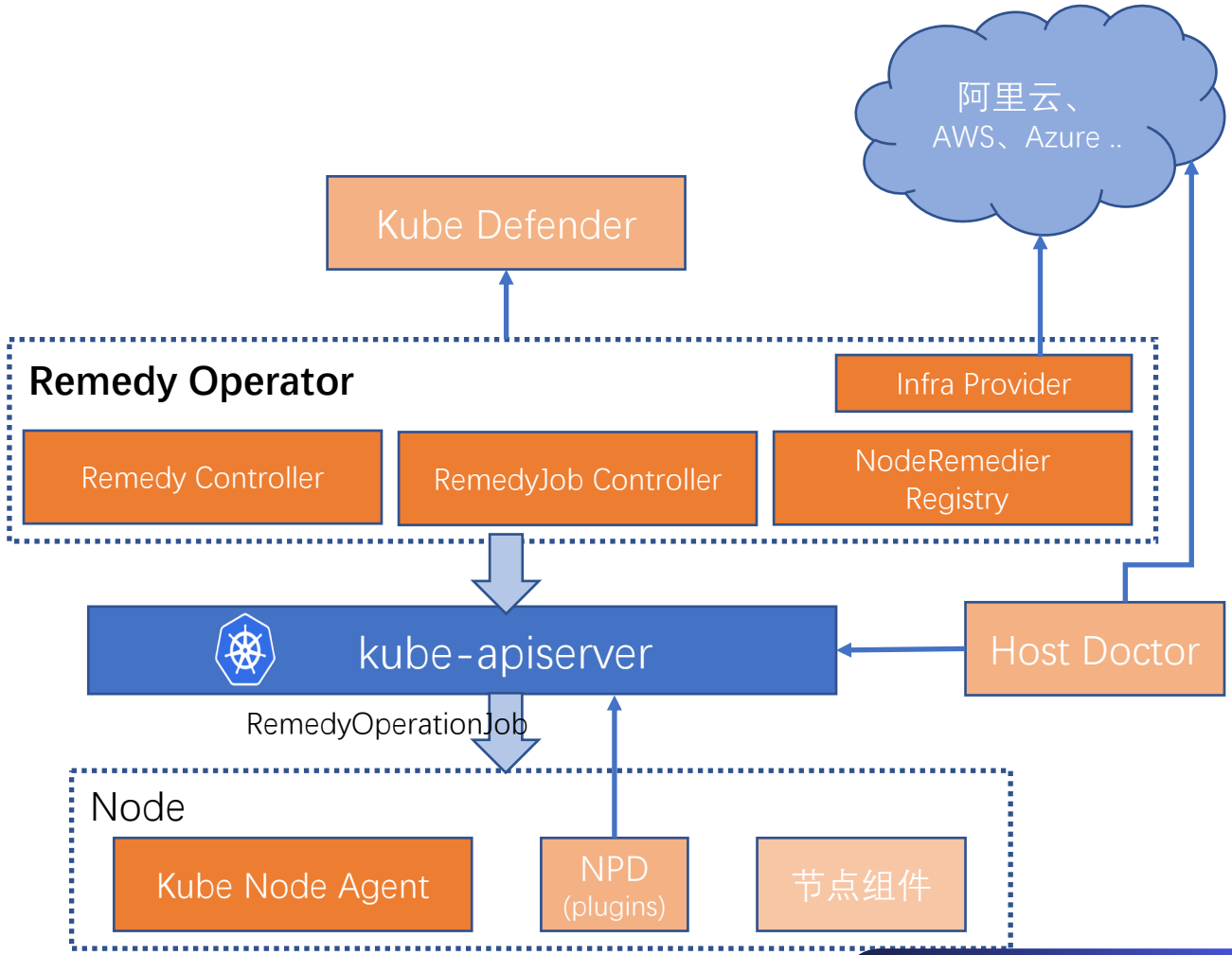
- ASIOps
  - ASI 组件变更统一 CD 平台
  - 上百集群 Pipeline 自动流水线发布
    - 测试 -> 预发 -> 正式
  - 变更后自动触发健康巡检

- KubeNode 组件升级
  - 逐批次灰度、暂停升级
  - 单机 watch 变化触发升级，高并发高效率
  - 健康巡检异常状态上报、暂停自动变更



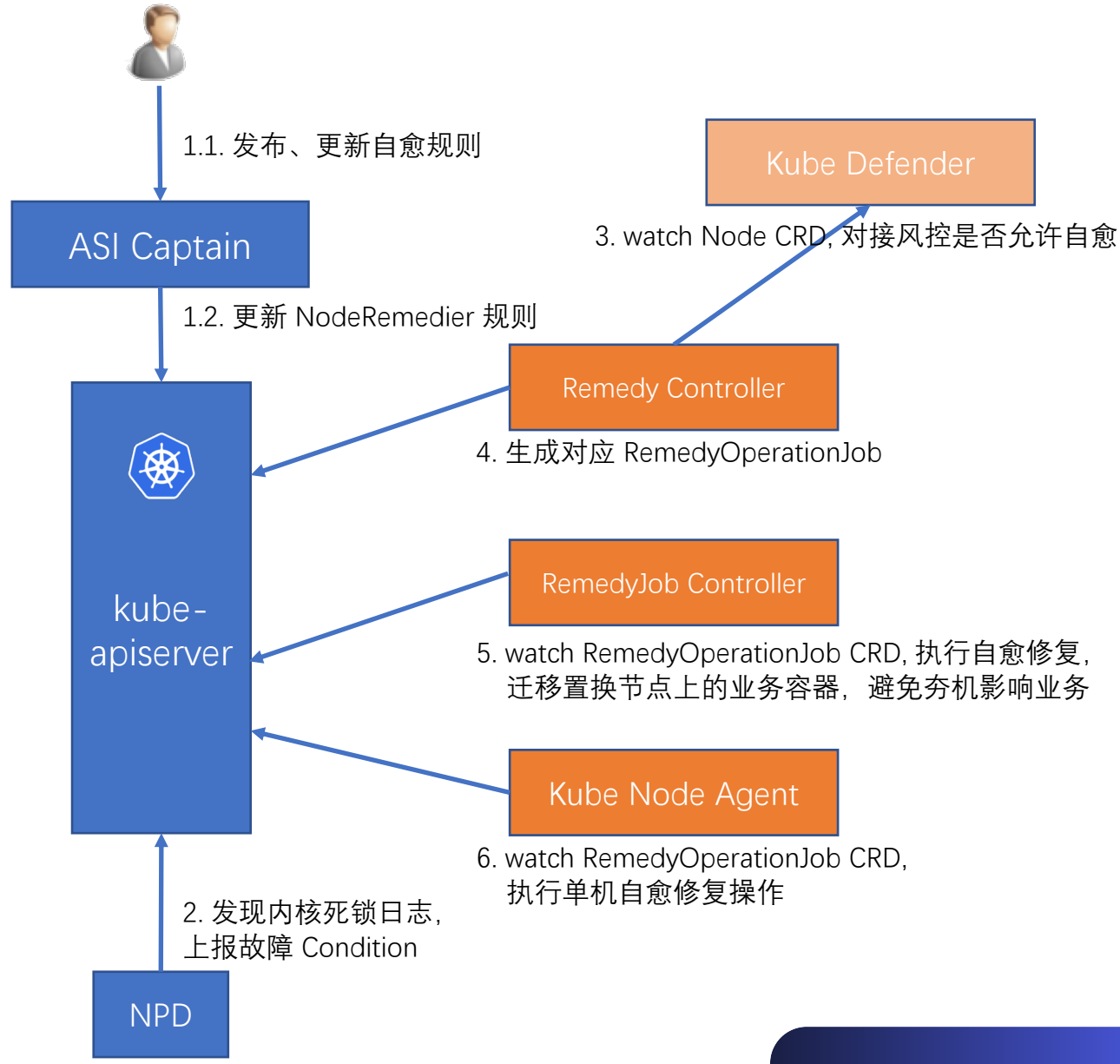
# KubeNode - Remedy Operator

- CRDs
  - NodeRemedier：节点故障修复规则
  - RemedyOperationJob：节点自愈修复任务
- Controllers
  - Remedy controller: 自愈控制
  - RemedyJob controller: 自愈任务控制
  - NodeRemedier Registry: 自愈规则注册中心
- Host Doctor: 中心故障诊断，对接主动运维事件
- NPD: 节点故障检测 (插件式: kernel/kubelet/docker/...)
- Kube Node Agent
  - 单机自愈修复任务执行



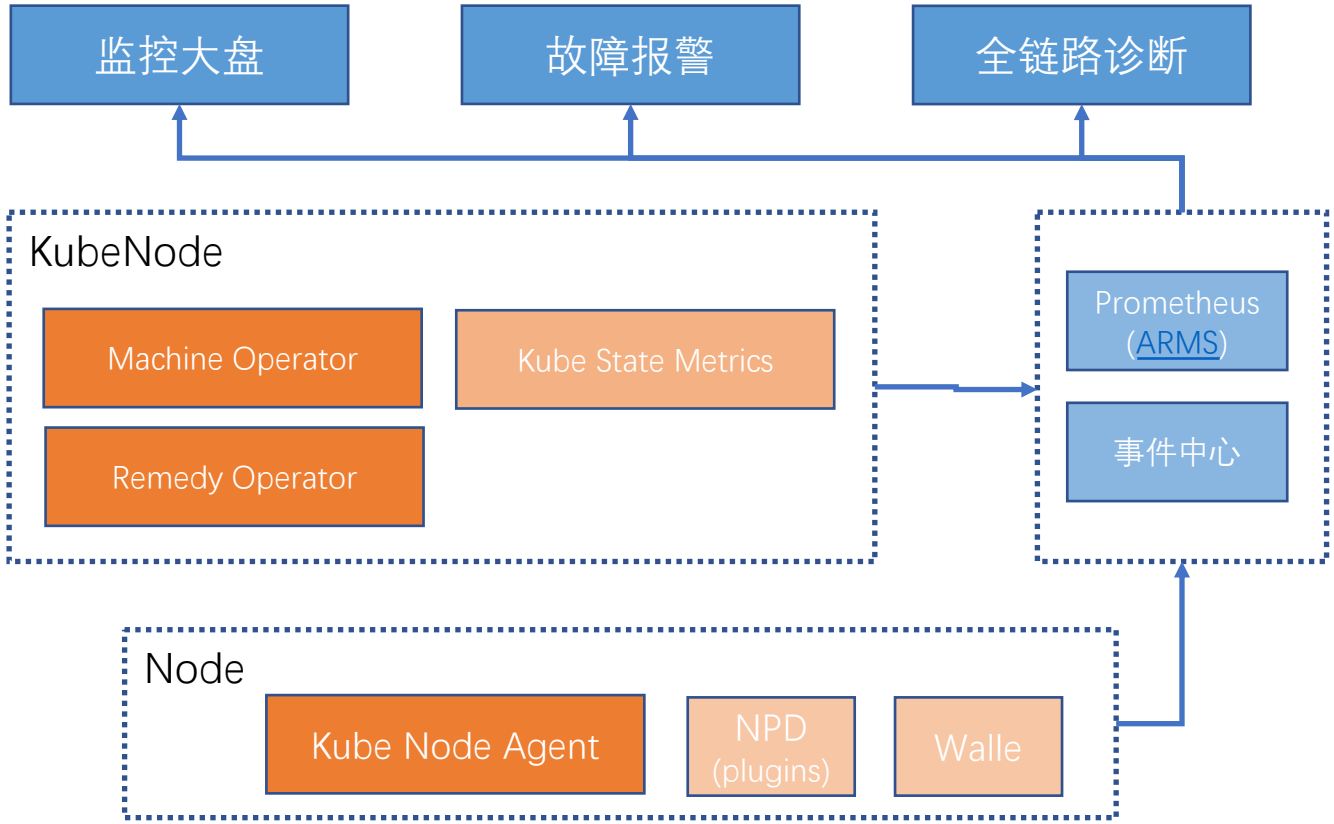
# Use Case: 夯机自愈

- 故障自愈
  - NPD -> Node Condition -> Remedy
- Remedy 自愈优势
  - 云原生自闭环自愈链路
  - 覆盖广：硬件、OS、组件
  - 秒级故障发现、分钟级故障自愈
  - 对接风控，防止自愈操作引发二次故障



# KubeNode 数据体系

- 数据体系
  - 数据采集链路
  - 统一数据采集和存储
- 数据平台应用
  - 资源利用率分析统计
  - 实时监控报警
  - 整体故障分析统计
  - 节点组件覆盖度、一致率分析
  - 节点自愈效率分析
  - 全链路诊断



# 未来展望

- 集团规模覆盖，支持经济体统一资源池节点管理
- 计划 2020 年下半年开源

# THANKS

首届 KubeCon 2020 线上峰...

22人

扫一扫群二维码，立刻加入该群。