



CLOUD NATIVE + OPEN SOURCE

Virtual Summit China 2020

Kubernetes Transforms Cloud Infrastructure for Local Life Services Giant

Guoliang Wang @Meituan-Dianping



自我介绍



王国梁

- 2017年加入美团点评，曾先后参与美团公有云、私有云以及HULK容器平台等调度系统研发和设计工作，目前主要负责Kubernetes集群运营和建设以及云原生技术的落地工作
- Kubernetes项目维护者和社区kube-scheduler的Reviewer之一

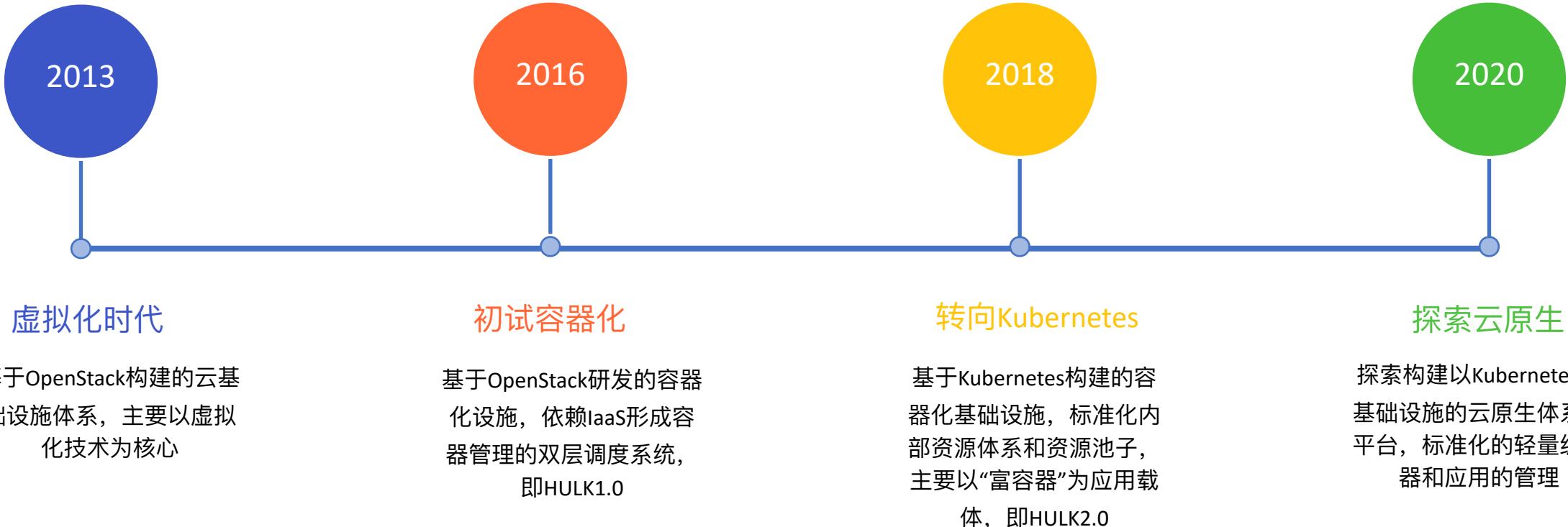


大纲

- 背景和架构
- 从OpenStack迁移到Kubernetes的障碍和收益
- 运营大规模Kubernetes集群的挑战和应对策略
- 总结和展望



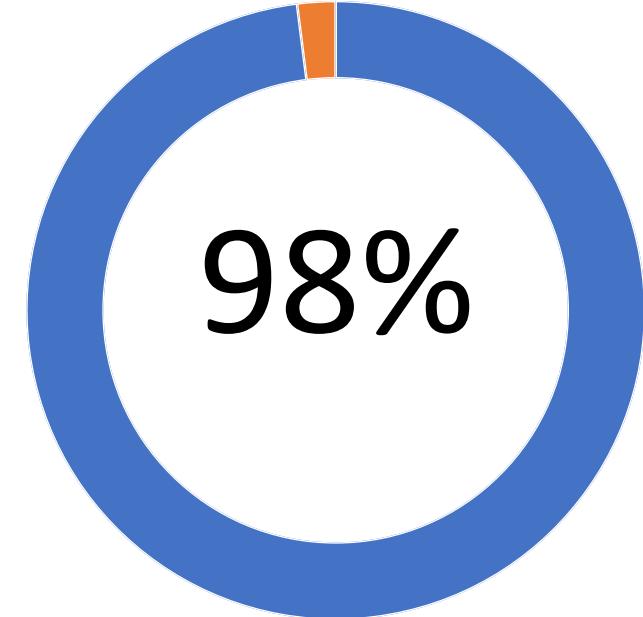
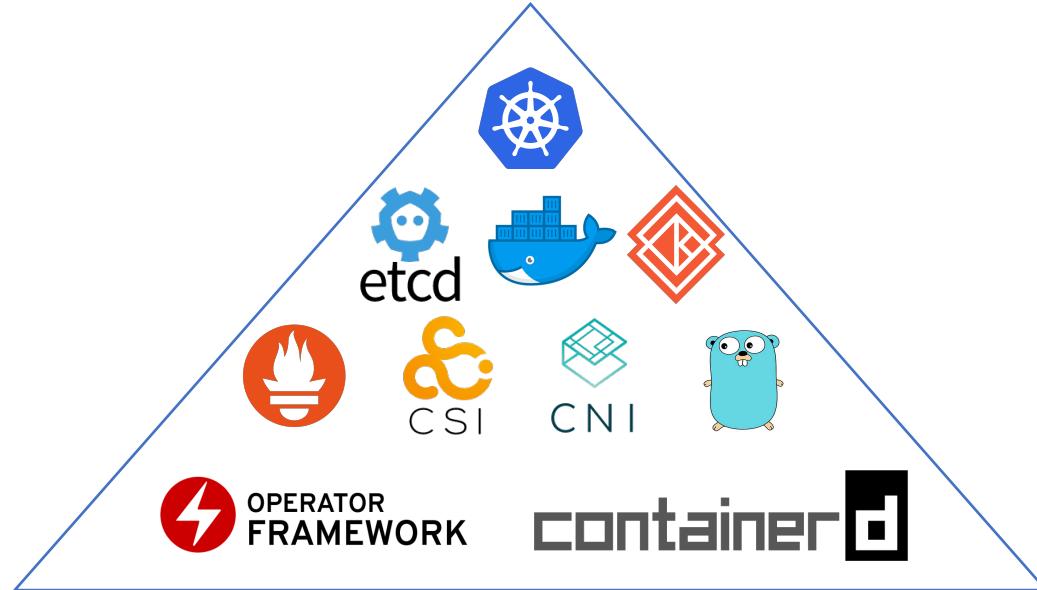
演变路线



美团点评基础设施现状



服务治理/DevOps/应用管理



- 10s Clusters
- 10,000s Nodes
- 100,000s Pods



统一资源调度架构

HULK

Serverless平台

服务网格

Squirrel

Databus

MySQL

Blade

Kubernetes APIServer

装机交付

配额管理

资源余量

资源画像

高可用/容灾

监控告警

调度策略

资源均衡

重调度

容量规划

CRI

CSI

CNI

Device Plugin

KubeNative-
云原生应用管
理引擎

故障自愈

集群巡检

MK8S-集
群管
理平台

ETCD集群

Docker

Kata

虚拟机

分布式存储

智能网卡

本地多盘

多机型兼容

私有网络

私有DNS

内核

私有云服务器

公有云服务器

镜像仓库

网络

存储

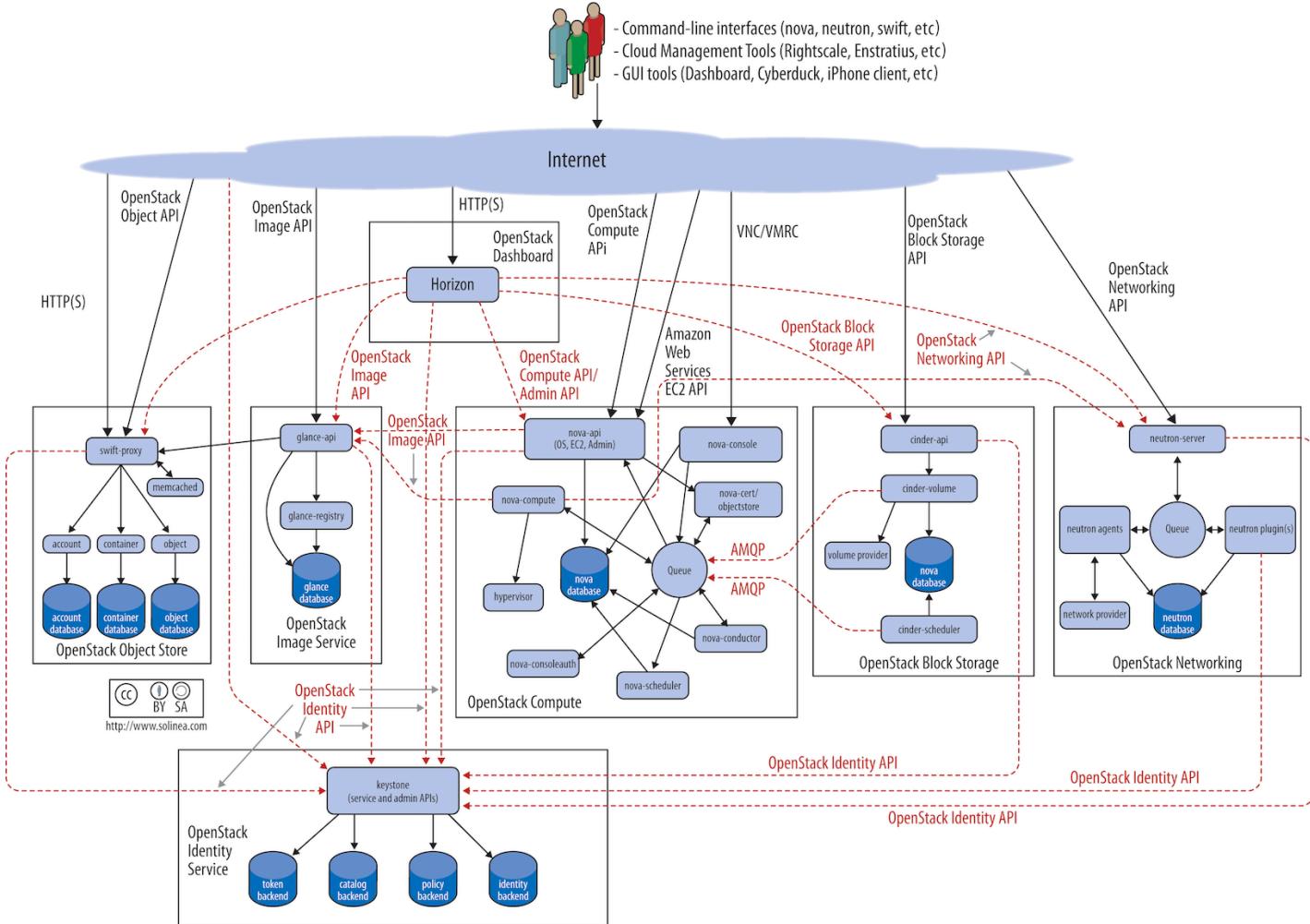
安全



大纲

- 背景和架构
- 从OpenStack迁移到Kubernetes的障碍和收益
- 运营大规模Kubernetes集群的挑战和应对策略
- 总结和展望

基于OpenStack的VM资源管理问题



Openstack Architecture

- 架构复杂，运营和维护困难
- 环境不一致性问题突出
- 虚拟化本身资源占用多
- 资源交付和回收周期长，不易灵活调配
- 高低峰明显，资源浪费严重

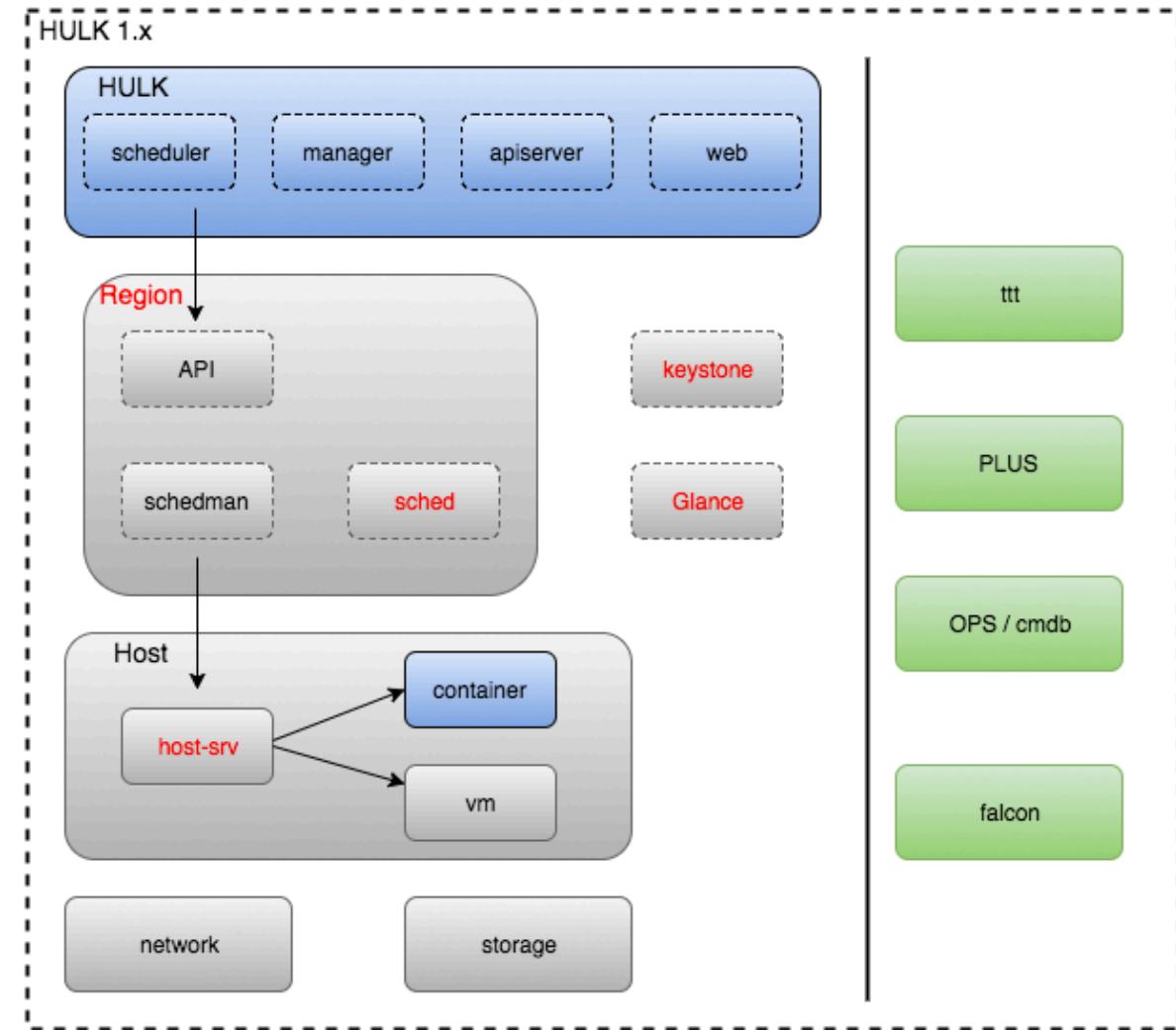
基于OpenStack的容器资源管理问题

- HULK1.0:

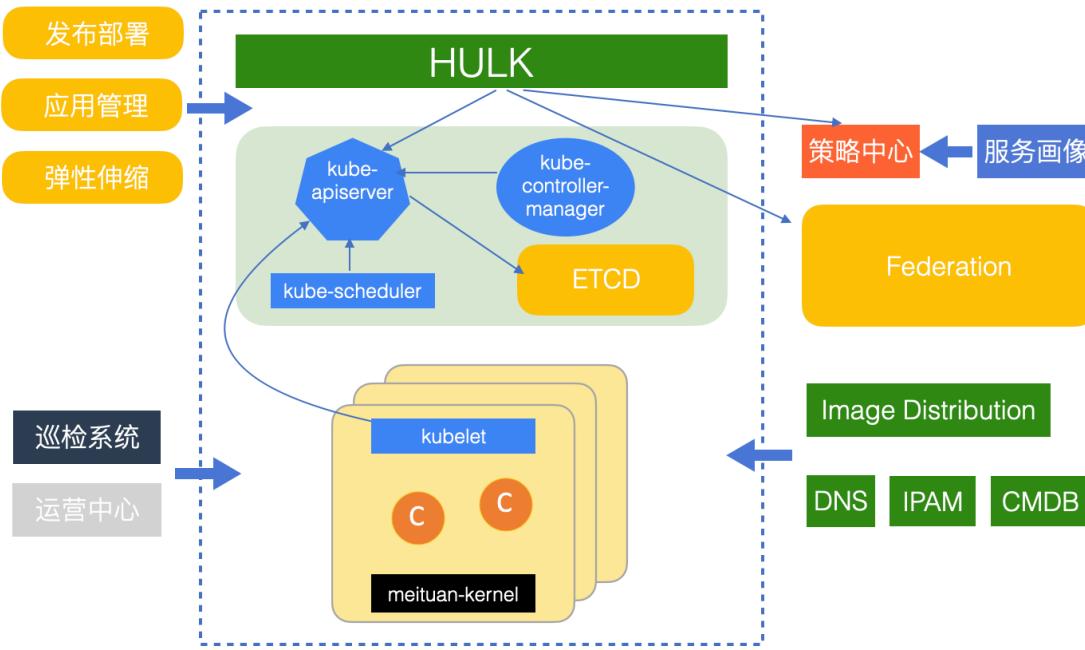
- 基于OpenStack实现,依赖IaaS层资源管理
- 构建了弹性伸缩平台, 做到业务高低峰资源弹性供给, 提升了资源效率
- 探索出可行的容器化道路和经验, 解决了环境一致性问题
- 资源交付和回收周期从分钟级到秒级

- 问题:

- **稳定性差**: 主要表现在与VM耦合共用底层资源引擎、双层调度、机房隔离性差
- **能力欠缺**: 故障节点迁移和恢复、资源类型单一、问题排查调用链长
- **扩展性**: 控制能力弱, 无法任意的扩展周边
- **性能**: 扩容周期20-40s; 容器隔离性和性能比VM低



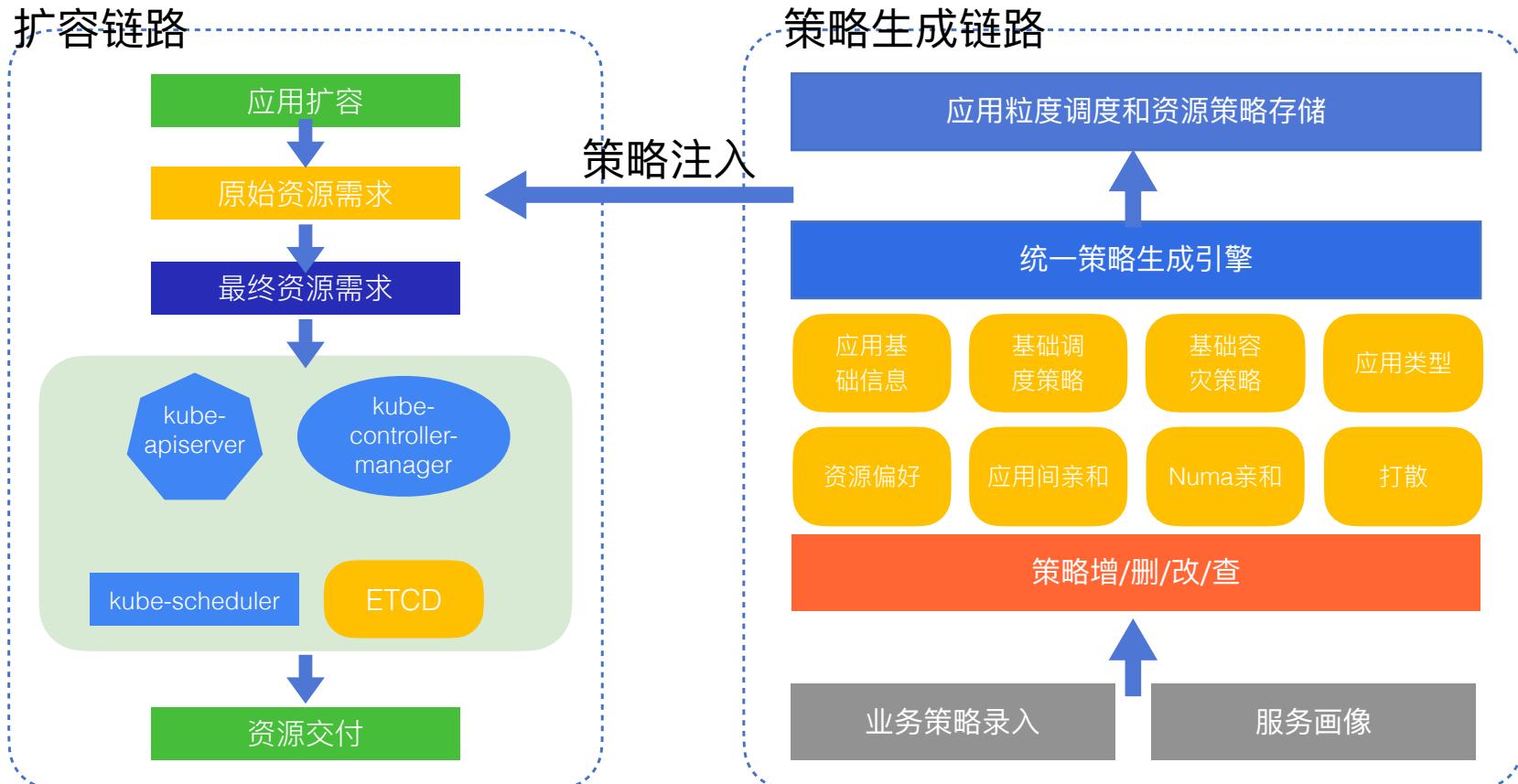
Kubernetes如何帮助我们改变？



- 基于Kubernetes构建的新的容器平台，完全兼容Kubernetes API
- 基于API的架构分层更清晰，扩缩容链路短
- 领域明确，依托于Kubernetes的强大的编排和管理能力，可独立迭代扩展

1. 复杂灵活、可配置的调度策略
2. 精细化资源调度和运营
3. 应用稳定性提升和治理
4. 平台业务的定制化转变
5. 业务资源优先级保障
6. 更进一步——云原生架构的落地

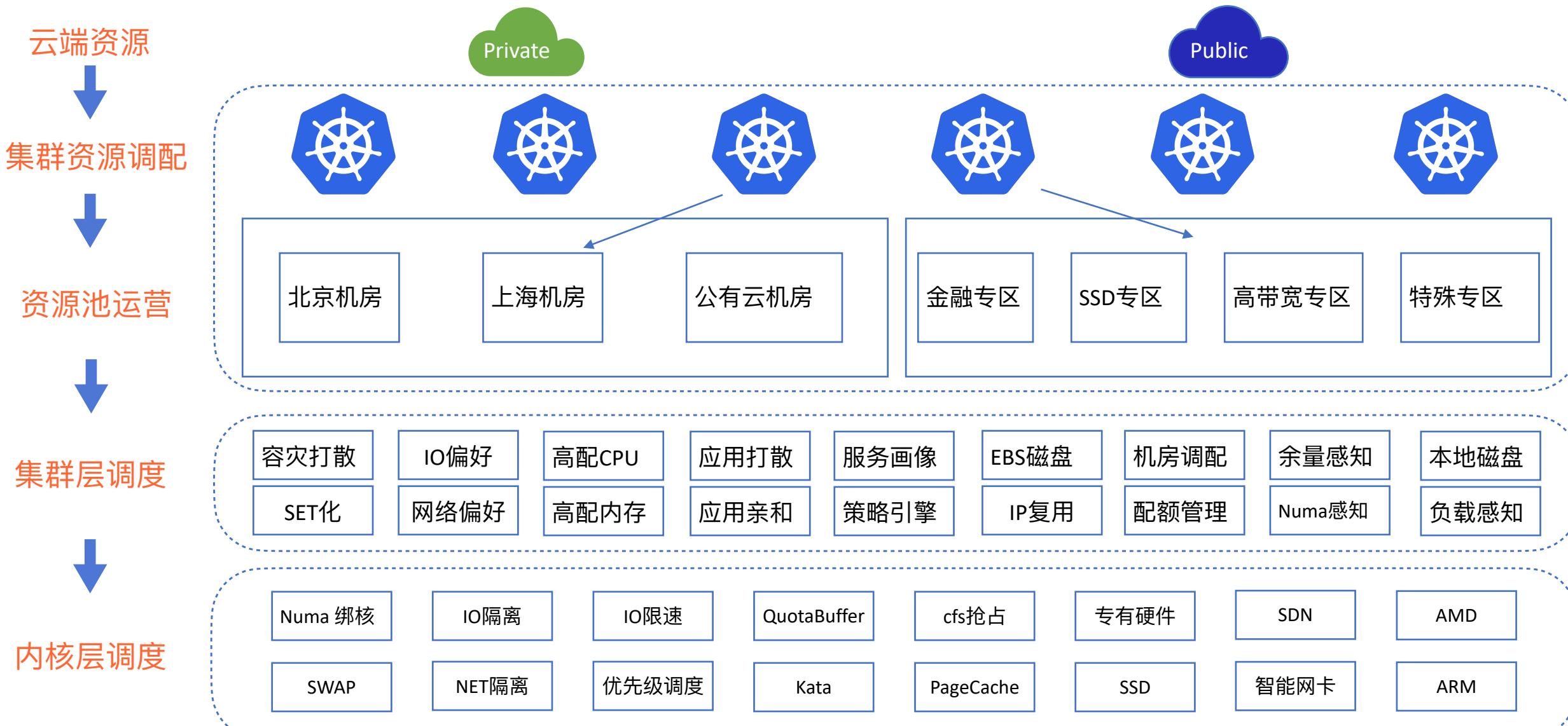
复杂灵活、可配置的调度策略



- 应用策略可配置、可变更、可实时生效
- 策略实时影响新扩容应用
- 服务画像自动跟进应用特征补充推荐策略



精细化资源调度和运营



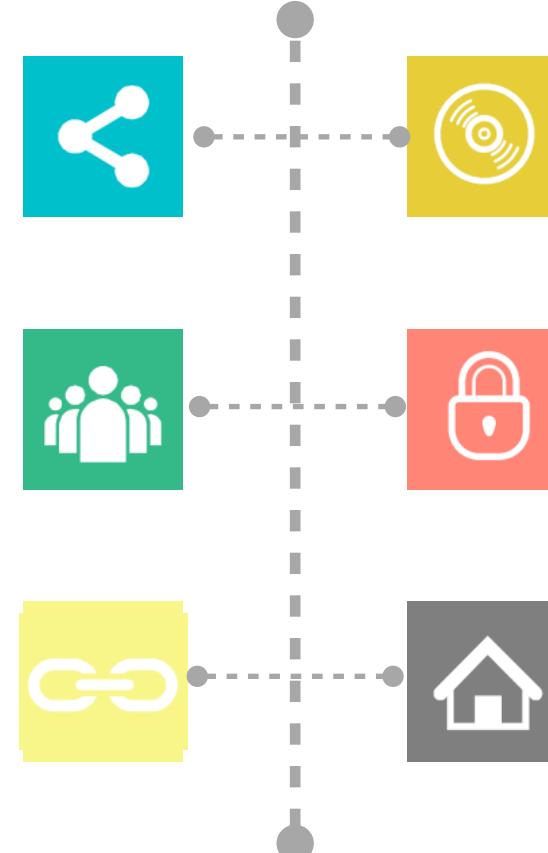


应用稳定性提升和治理

容器复用，避免宿主机重启导致容器数据丢失

负载感知，优先调度到空闲机器，减少资源竞争

故障容器告警和自动处理，先于业务发现异常



Numa Node的感知和绑定，避免跨Node的CPU访问，减少抖动

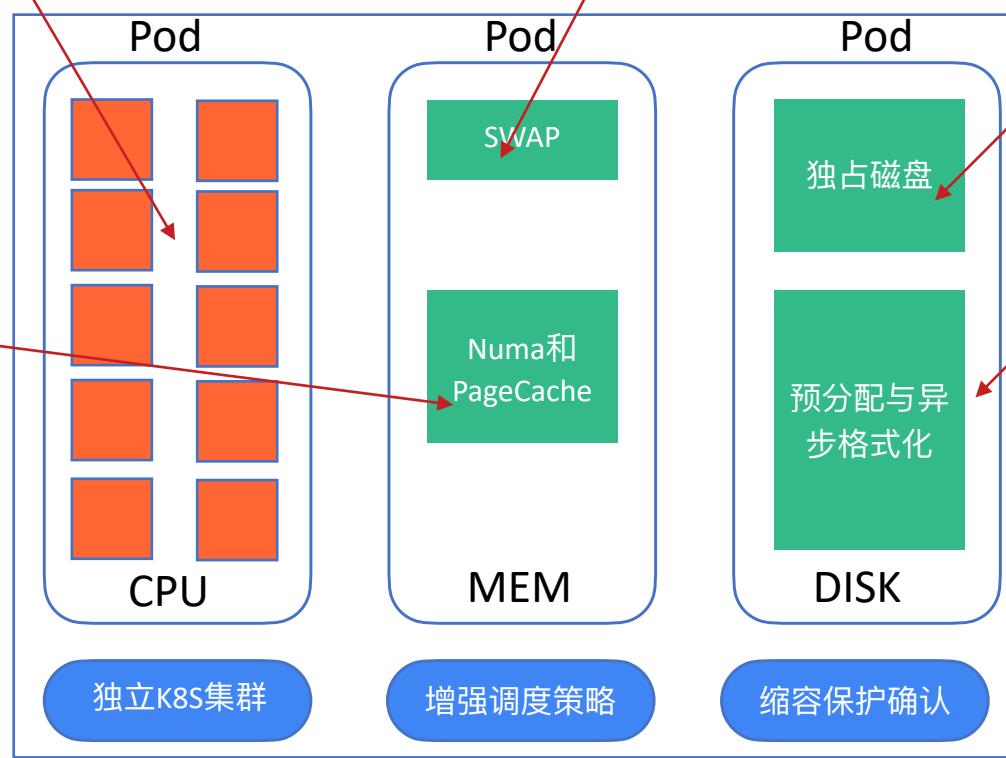
基于服务画像和服务特征的相同类型资源偏好的应用打散

特殊资源应用专区隔离部署



平台业务的定制化转变案例

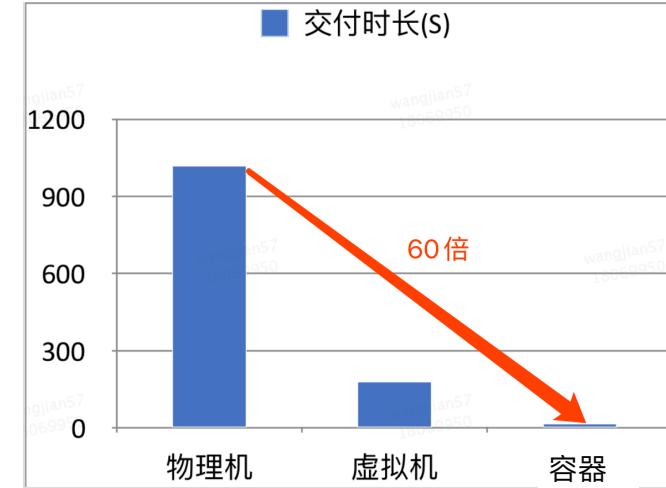
独占绑定CPU Cores



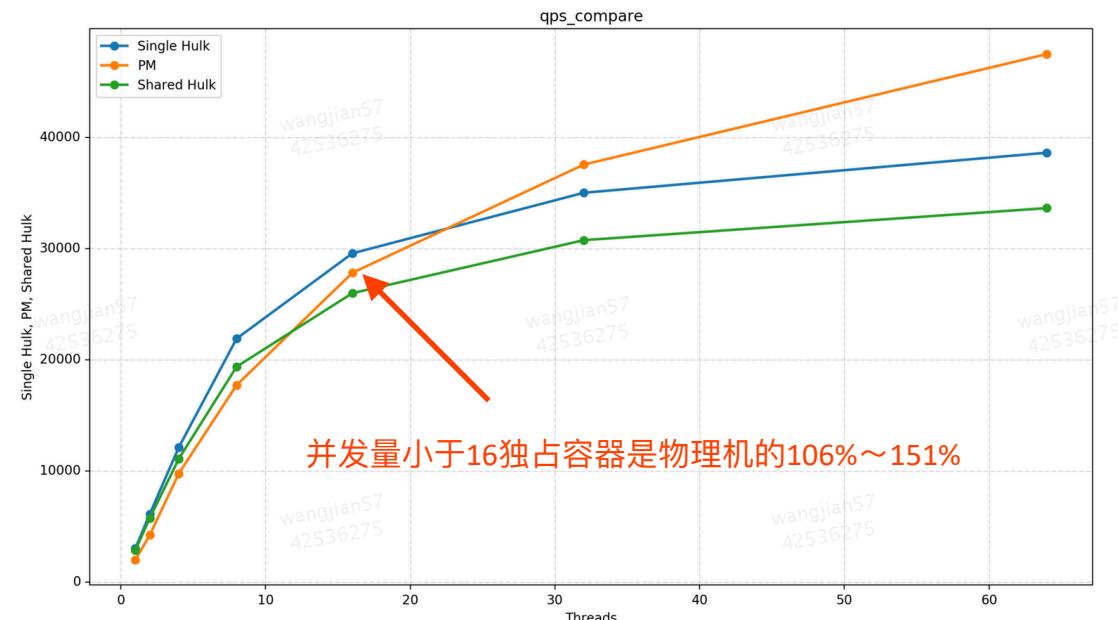
定制SWAP大小应对高流量

磁盘IOPS隔离

提升扩缩容效率



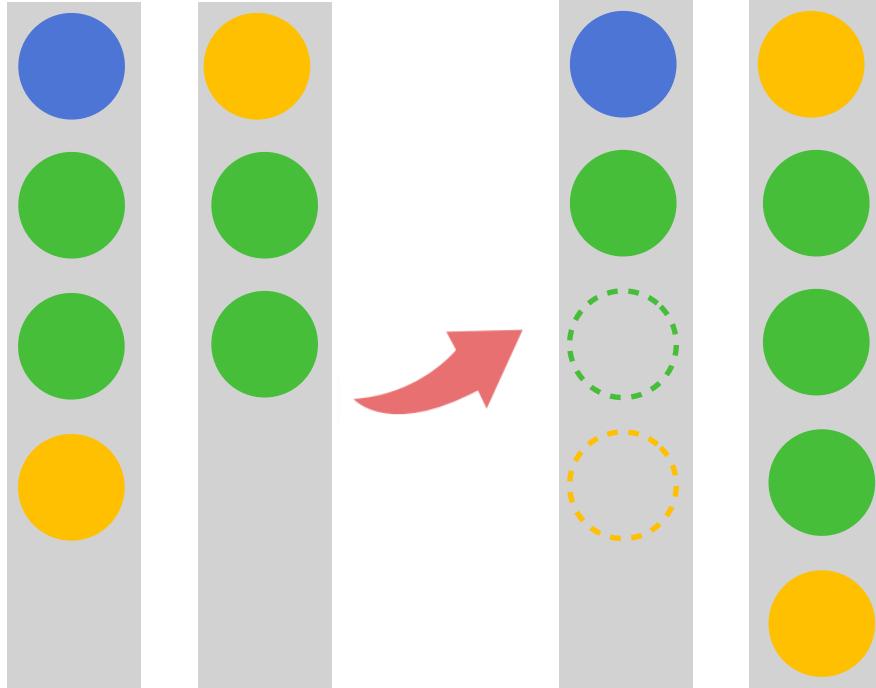
并发量小于16独占容器是物理机的106%~151%



收益：规模效应、收益明显、效率提升高



业务资源优先级保障



业务资源配额/专区资源，按预算分配保障



弹性资源池/公有云资源应对突发资源需求



按应用等级/应用类型优先级调度



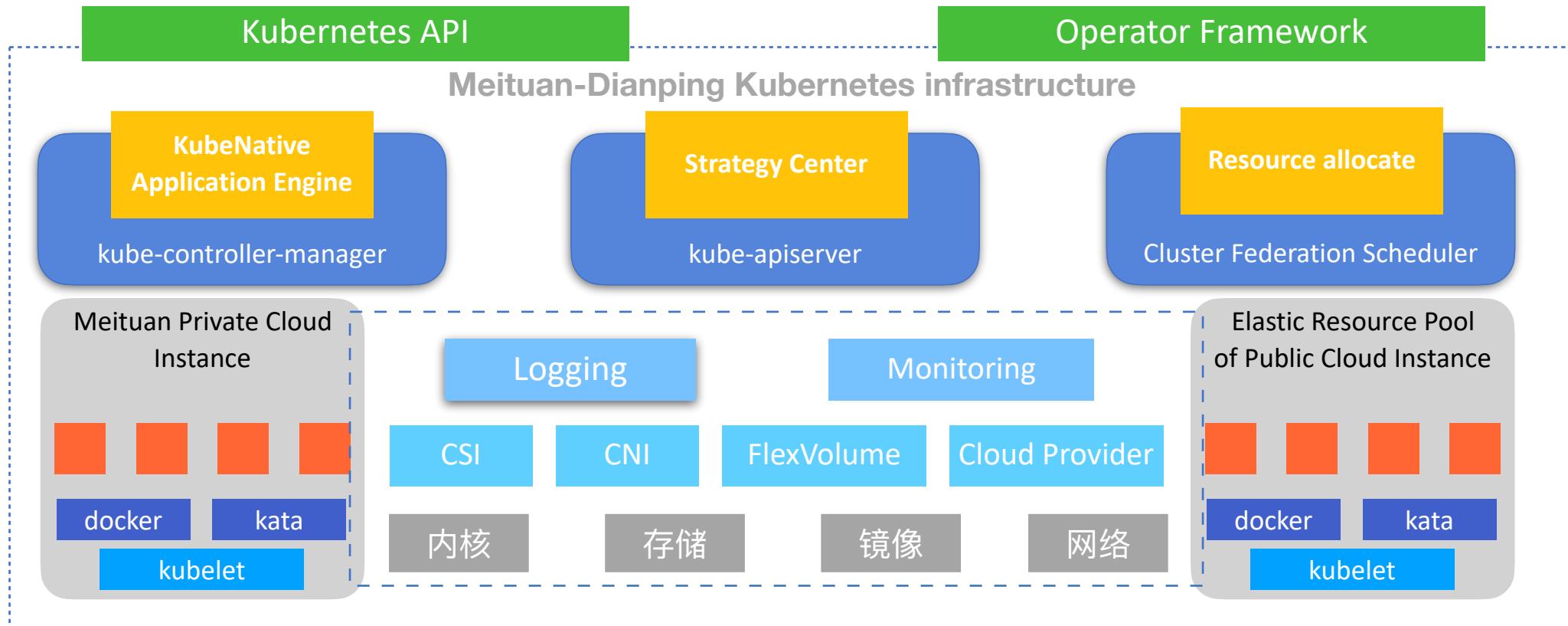
多集群/机房资源保障突发故障

更进一步——云原生架构的落地

Cluster Mgmt System
(MK8S,Monitor/Alaert System,Resource Dashboard)

HULK/Elastic scaling system/
PaaS

Serverless,TiDB,listio





我们的收益

- 完成了全公司业务98%的容器化，大大提升资源管理的效率和业务稳定性
- Kubernetes可用性99.99+%
- Kubernetes成为美团点评集群管理平台的内部标准



大纲

- 背景和架构
- 从OpenStack迁移到Kubernetes的障碍和收益
- 运营大规模Kubernetes集群的挑战和应对策略
- 总结和展望

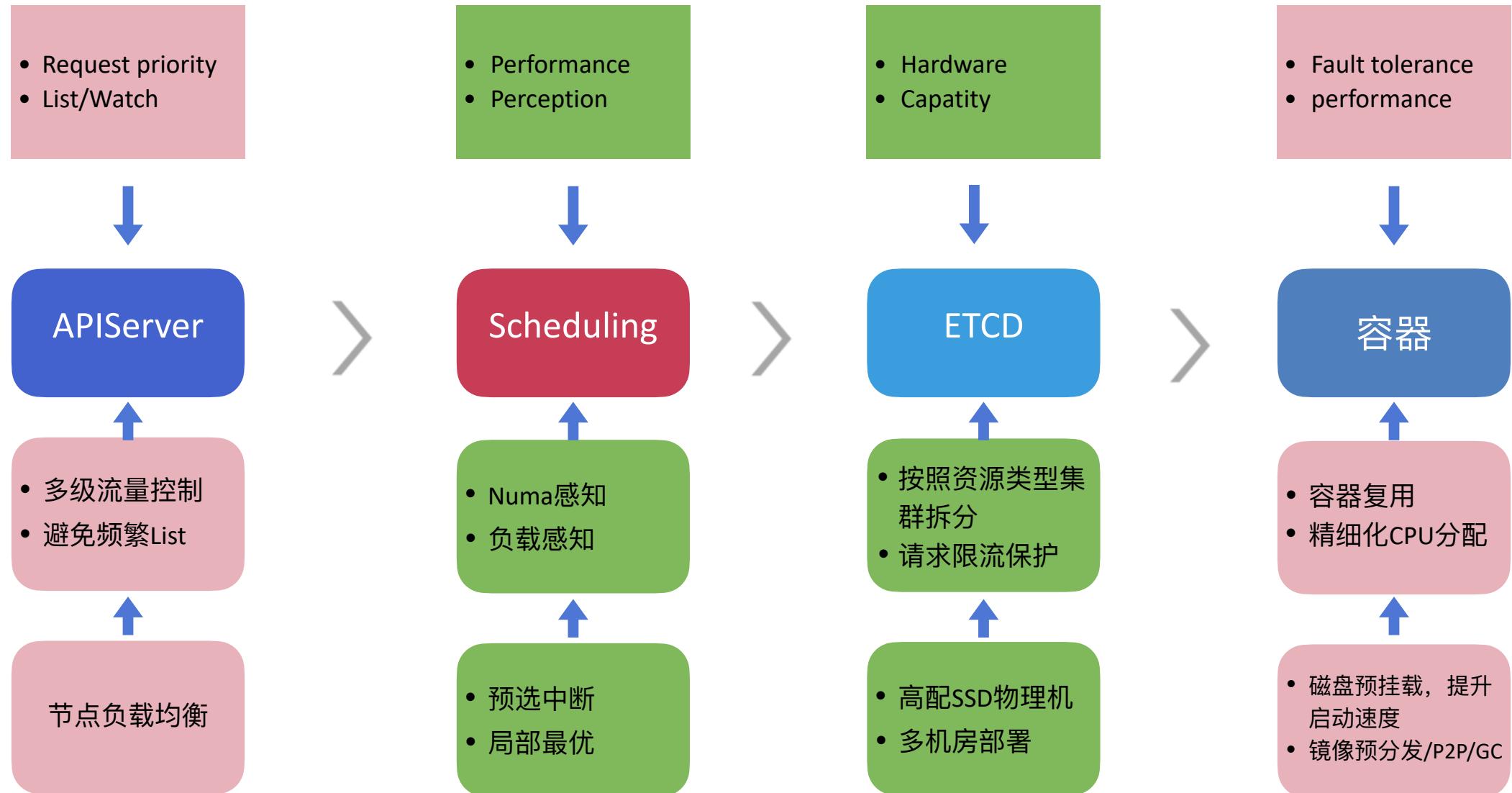


集群核心组件优化与升级

- 集群规模大且增长快速，可用性和性能要求高
 - 各个组件我们需要关注什么？
- 如何进行集群升级
 - 集群规模大，旧版本稳定性也低，如何确保只能成功不能失败？
 - 如何保证资源不变和业务无感知的情况下稳定可靠地进行集群升级？



核心组件分析与优化

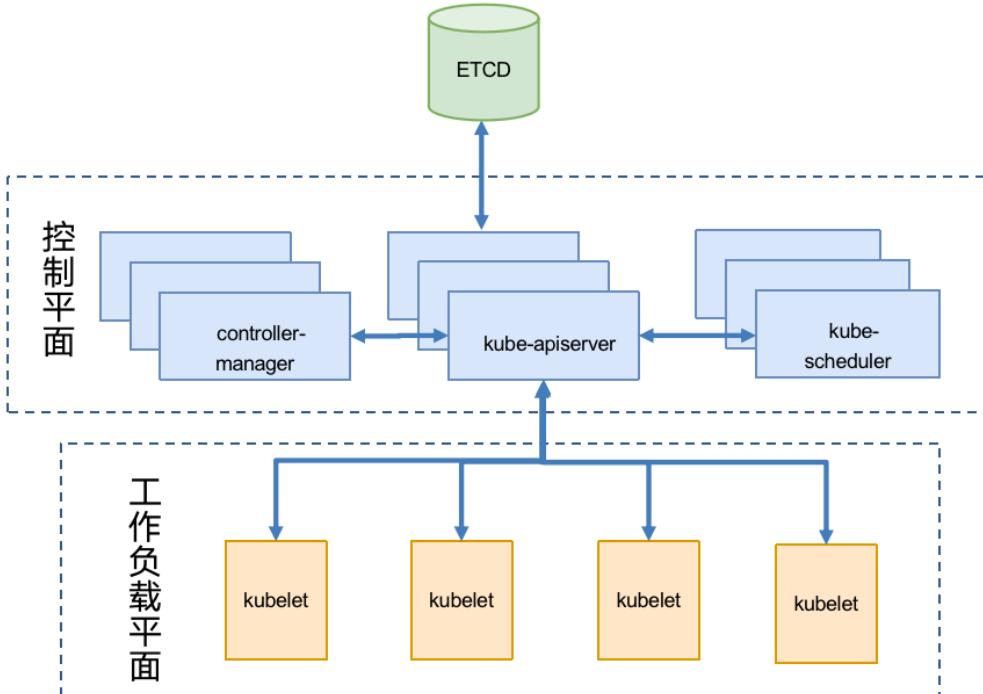




让升级不再困难

业界方案的问题：

- 可升级版本有限制，无法直接跨大版本升级
- 控制平面升级风险可控性差、灰度能力弱
- 用户有感知，容器需要新建，成本高

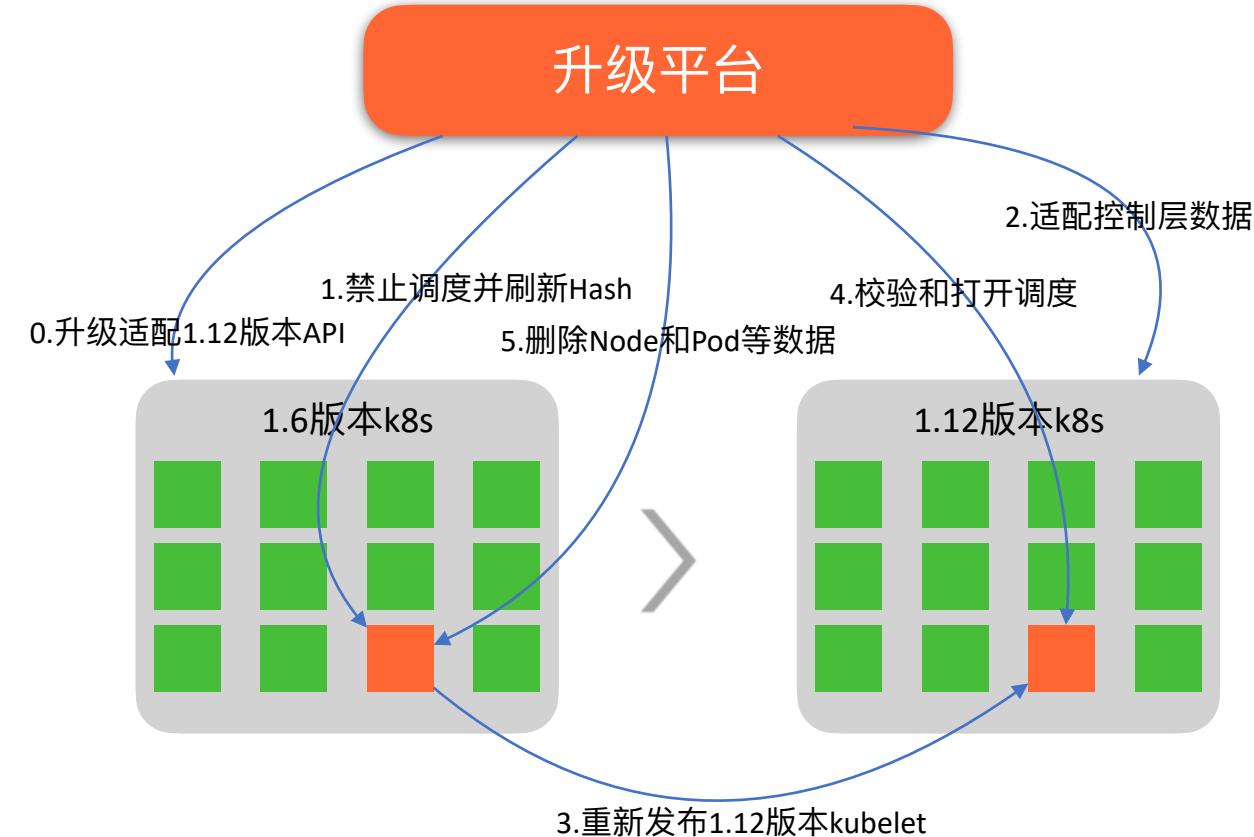




让升级不再困难

方案特点：

- 大规模生产环境的集群升级不再是难题
- 解决了现有技术方案风险可控能力差的问题，风险降到宿主机级别，升级更安全
- 通用性强，可做到任意版本升级，且方案生命周期较长
- 优雅解决了升级过程中容器新建问题，真正做到了原地热升级

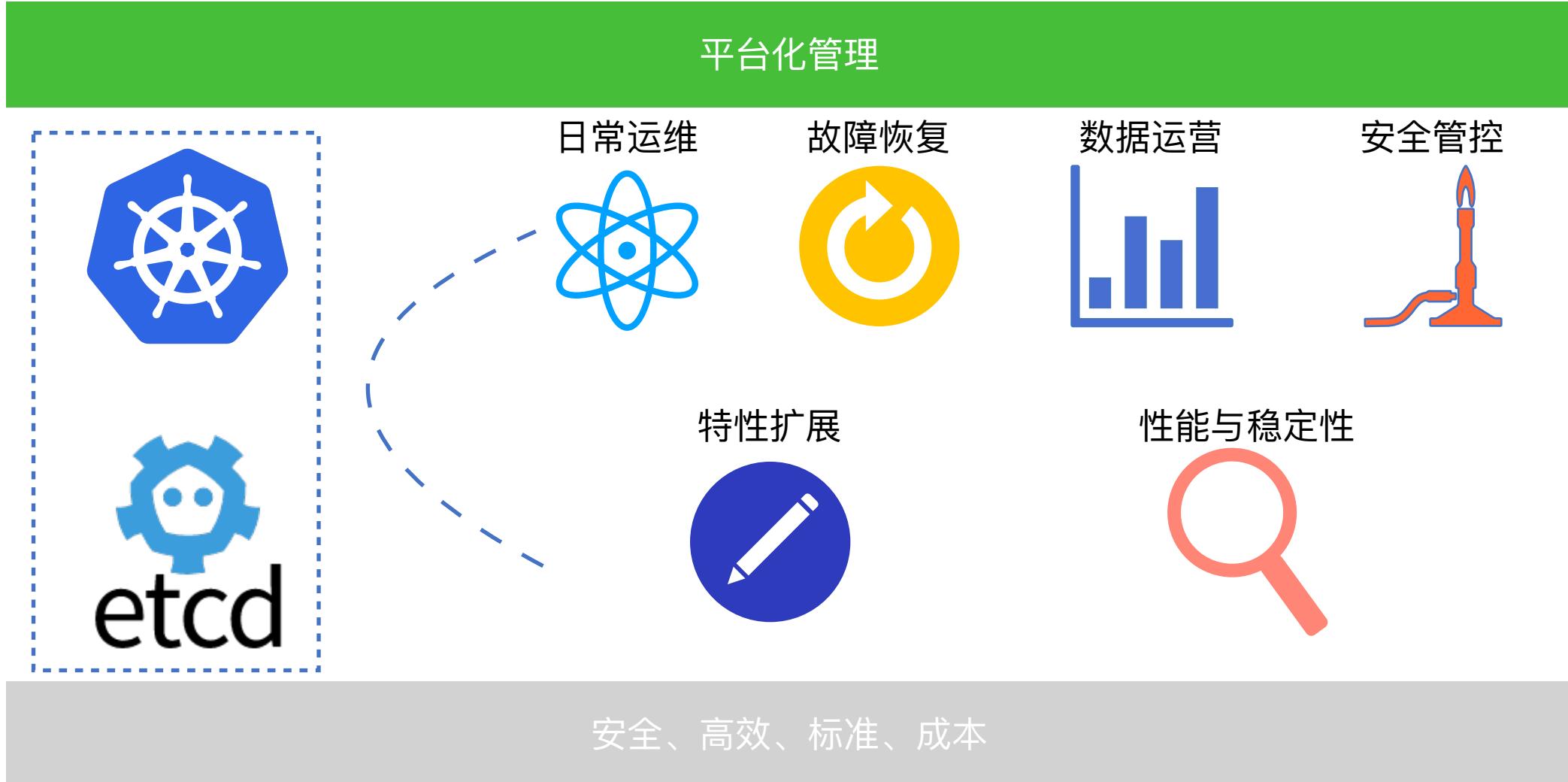




能力建设和运营效率

- 多维度的能力建设
 - 我们都需要从哪些方面入手?
 - 在人力有限情况下保障运营质量
 - 集群管理平台化
 - 运维处理流程化
 - 故障处理自动化
 - 集群运营智能化

不同维度的能力建设





运营效率和工具化

平台化

- 标准化集群生命周期管理
- 可视化运维管理，避免命令行操作

流程化

- 大大提升了运维效率——例如集群搭建周期从天级降低为分钟级
- 避免运维多次切换平台
- 降低了误操作的可能性

运营效率

- 告警自愈——每周人均告警降低40%/告警处理耗时从分钟级降低为秒级/运维工作量降低90%
- 自动巡检——周期性集群巡检，补全了问题盲点/风险感知为天粒度

自动化

- 精细化调度
- 故障分析和预测
- 容量预测

智能化

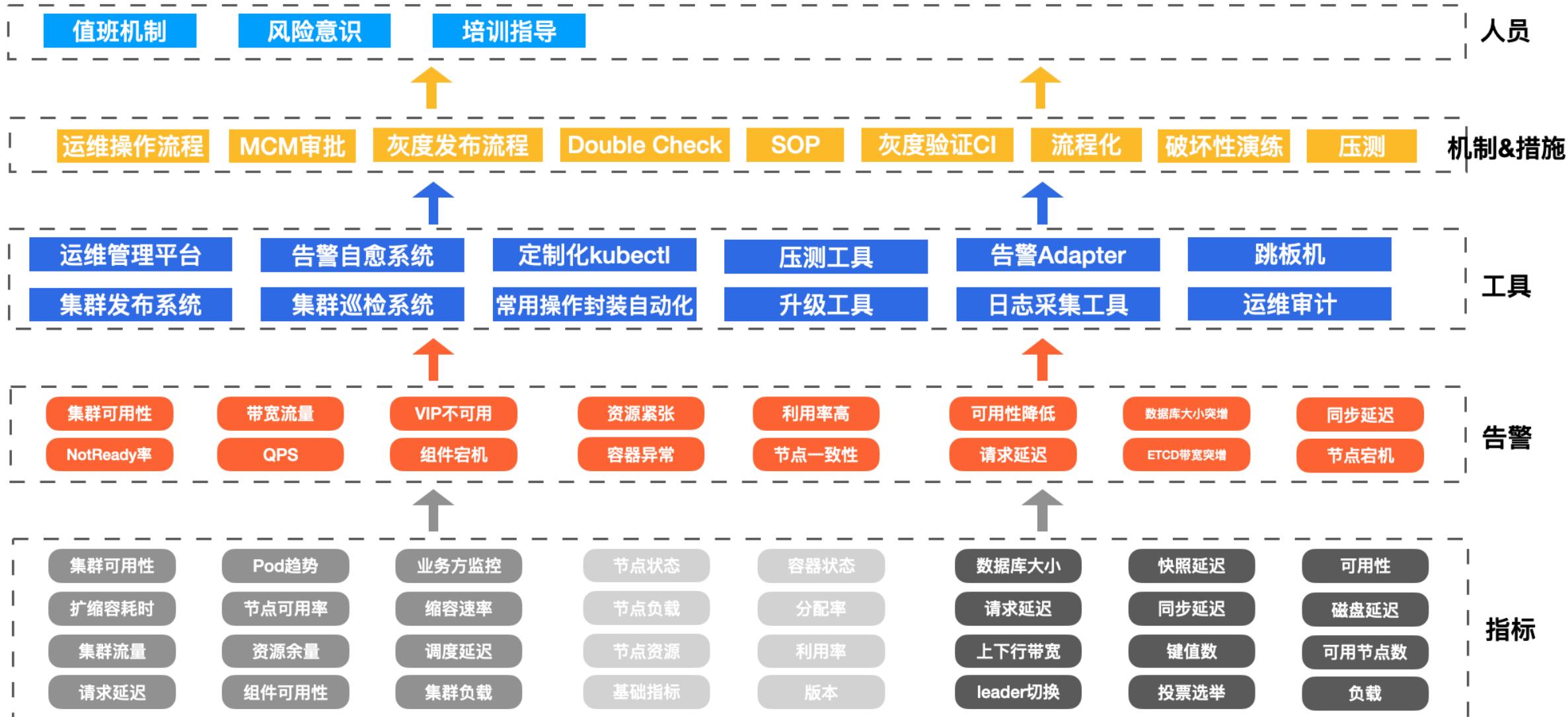


风险运营和可靠性保障

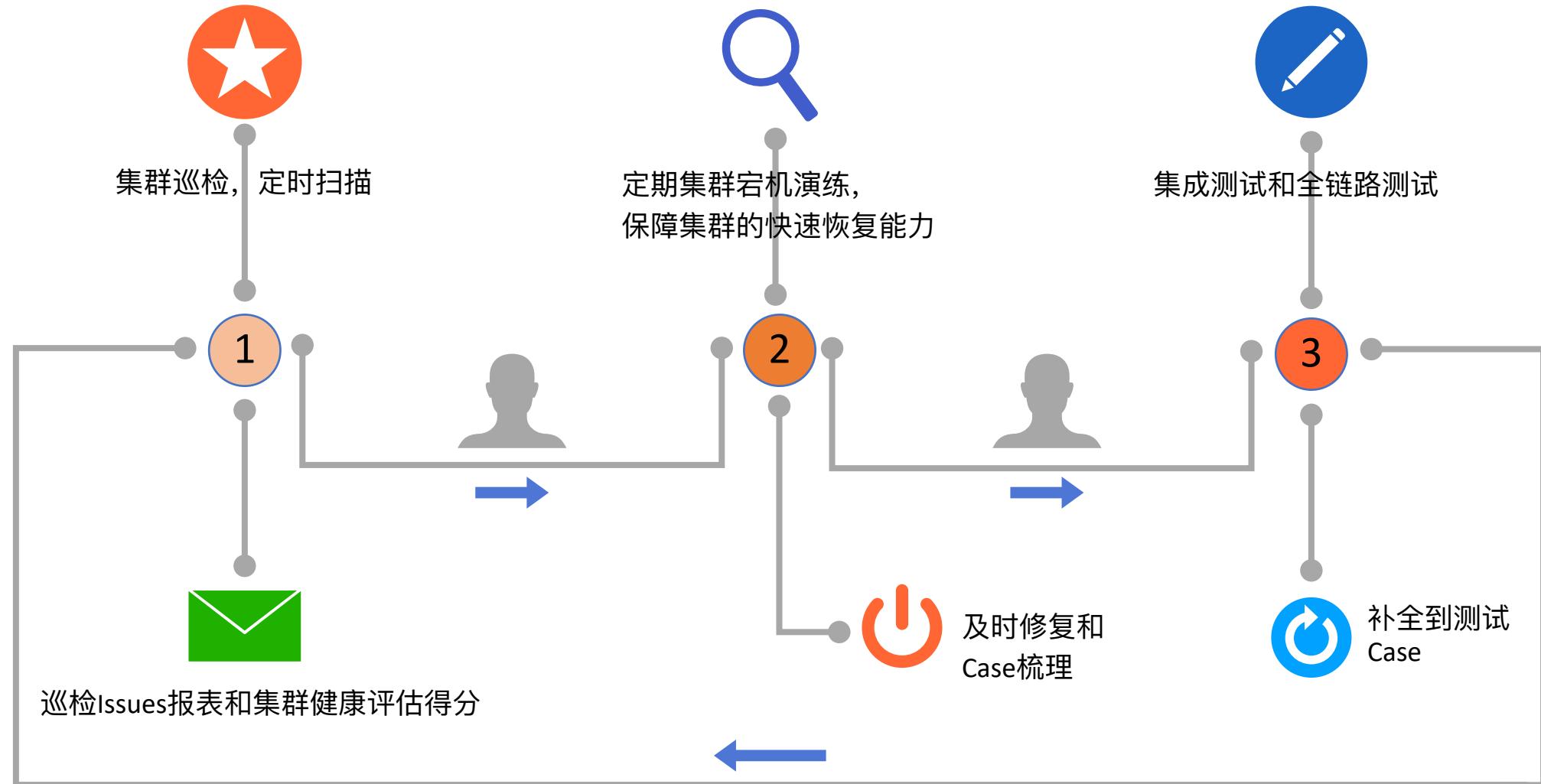
- 风险控制
 - 运维压力大、复杂度高
 - 如何避免运维故障和风险？
- 可靠性保障
 - 集群巡检与演练
 - 全链路测试



风险控制策略



可靠性保障策略





大纲

- 背景和架构
- 从OpenStack迁移到Kubernetes的障碍和收益
- 运营大规模Kubernetes集群的挑战和应对策略
- 总结和展望



经验总结

Kubernetes API
第一原则

把握节奏
有序升级

落地要以解决
用户痛点为突
破口

价值展现



未来展望

- **统一调度**: VM会长期少量存在，基于Kubernetes统一管理虚拟机和容器
- **VPA**: 垂直 Pod 自动扩缩，提升资源效率
- **云原生应用管理**: 探索云原生应用管理的落地，降低业务运维复杂度，提升应用管理能力和效率，让业务聚焦于业务本身；
- **云原生架构落地**: 推进各中间件、大数据、搜索等PaaS平台的云原生系统落地



Thanks

Q&A



我的微信号：Apply-A-Pod



CLOUD NATIVE + OPEN SOURCE

Virtual Summit China 2020

招聘：基础架构研发/Kubernetes相关岗位

邮箱：wangguoliang#meituan.com



美团点评