

# Sigma调度&Pouch容器创新实践

叔同

---

系统软件事业部 打造具备全球竞争力、效率最优的系统软件

# 个人介绍

- 丁宇，阿里花名叔同
- 天猫双11技术大队长，资深技术专家
- 2010年加入淘宝网、8次参与双11作战
- 阿里高可用架构负责人、双11稳定性负责人
- 阿里容器、调度、集群管理、运维技术负责人
- 推动和参与了双11几代技术架构的演进和升级



# Agenda

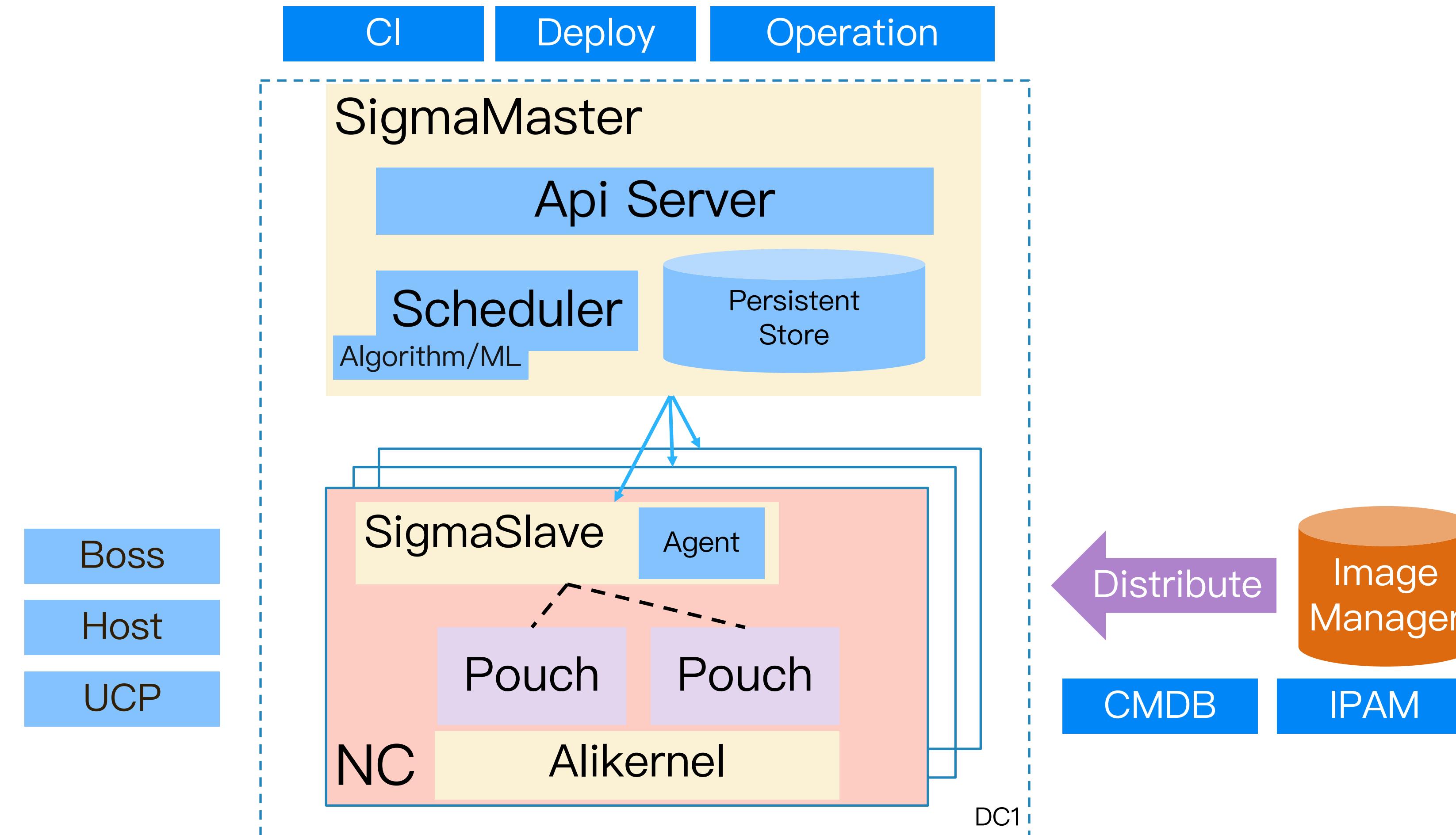
01 Sigma调度

02 混合部署

03 Pouch容器

04 未来路线

# Sigma调度&集群管理体系



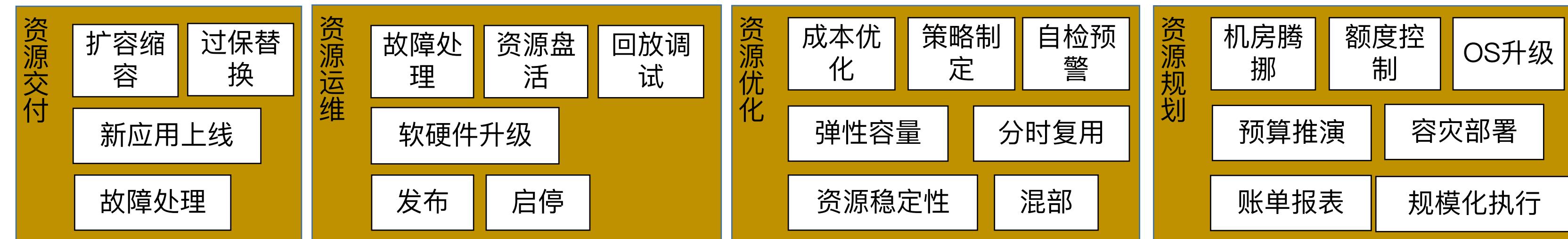
- 始于2011年，以调度为中心的集群管理体系
- 面向终态的架构设计；三层大脑合作联动管理
- Go语言重构，17年兼容Kubernetes API，和开源社区共同发展

# Sigma业务架构

## 业务领域



## 业务场景



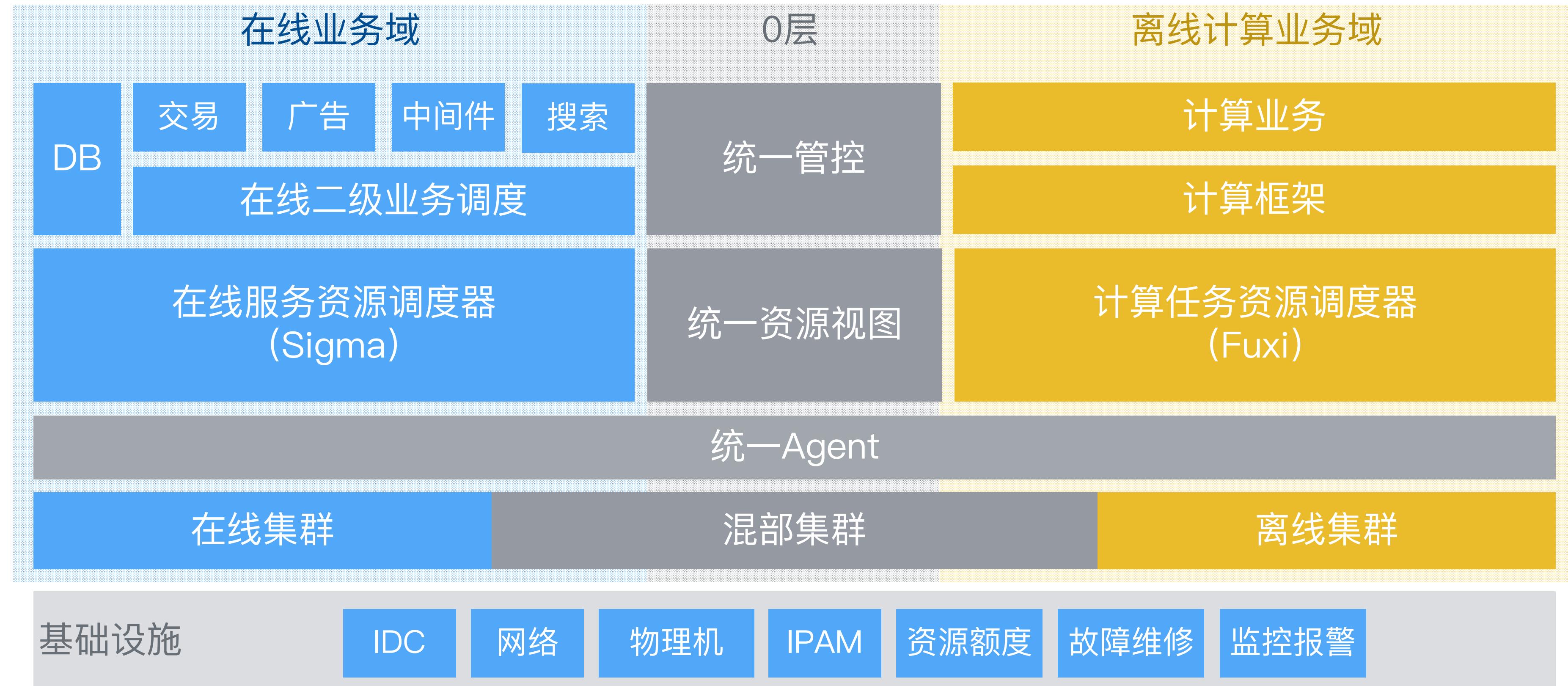
## 调度能力



## 基础设施

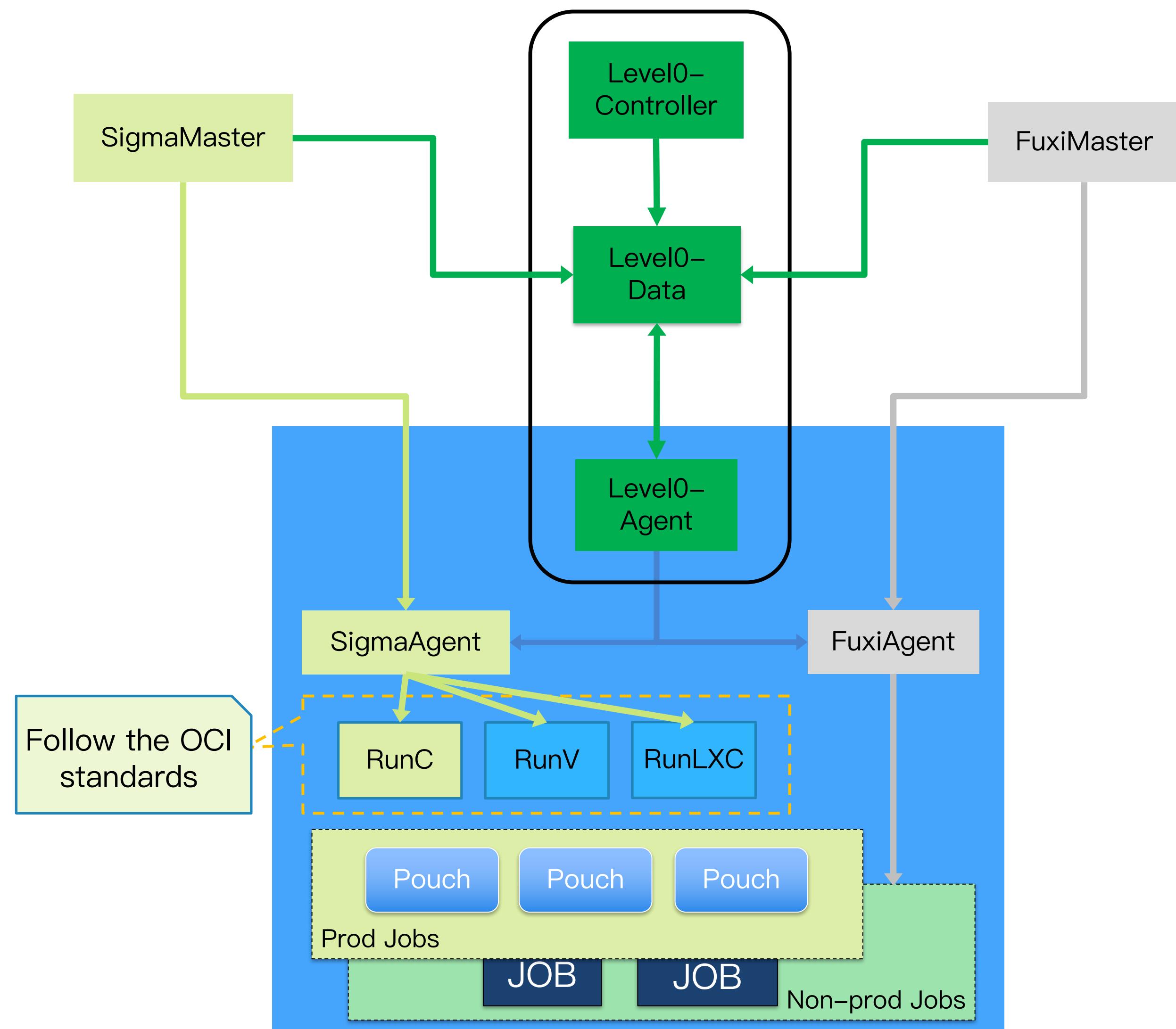
IDC建设  
网络架构  
物理机交付  
IP管理  
网络变更  
硬件采购  
物理机额度  
故障维修

# 调度系统现状



- 合并资源池，提升在线率、分配率去Buffer，空间维度优化
- 弹性分时复用，时间维度优化，共节省超过5%的服务器资源
- 发挥了统一调度、集中化管理的优势，释放规模效益下的红利

# Sigma与Fuxi混部架构



- 在线服务生命周期长/定制策略复杂/时延敏感；计算任务生命周期短/大并发高吞吐/时延不敏感
- 通过Sigma和Fuxi完成在线服务、计算任务各自的调度，计算共享超卖
- 通过零层相互协调资源配比做混部决策，通过内核解决资源竞争隔离问题
- 架构非常灵活，一层之间共享状态调度，一层之上定制二层调度
- 阿里混部始于2014年，已大规模铺开

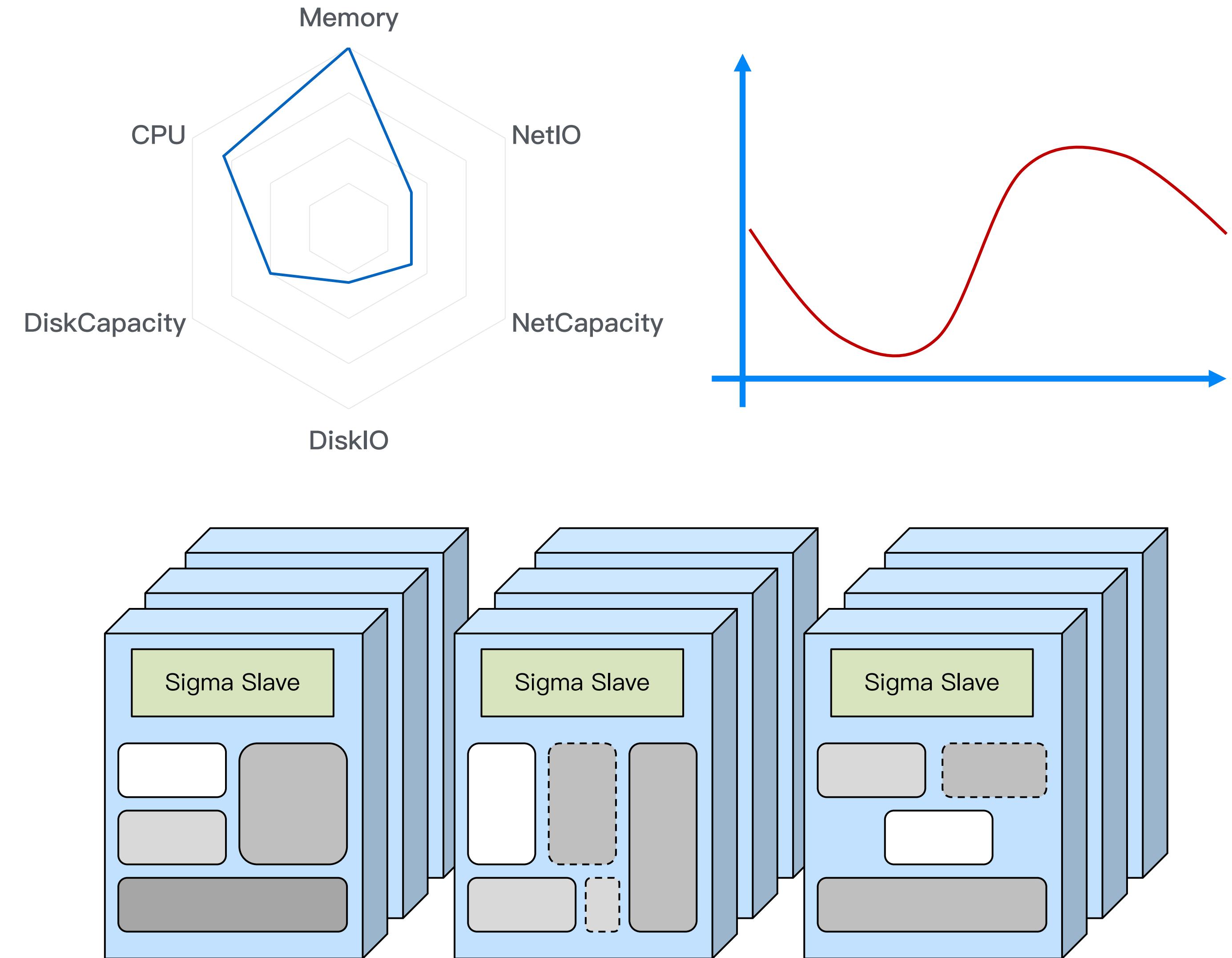
## 内核资源隔离

- CPU HT资源隔离：Noise Clean内核特性，解决超线程资源争抢问题
- CPU 调度隔离：CFS基础上增加Task Preempt特性，提高在线服务调度优先级
- CPU 缓存隔离：CAT，三级缓存(LLC)通道隔离(Broadwell及以上)
- 内存隔离：CGroup隔离/OOM优先级；Bandwidth Control实现带宽隔离
- 内存弹性：在线闲置时计算突破memcg limit；在线需要内存时计算及时释放
- 网络QoS隔离：TC增强，管控金牌；在线银牌；计算铜牌，分级保障带宽

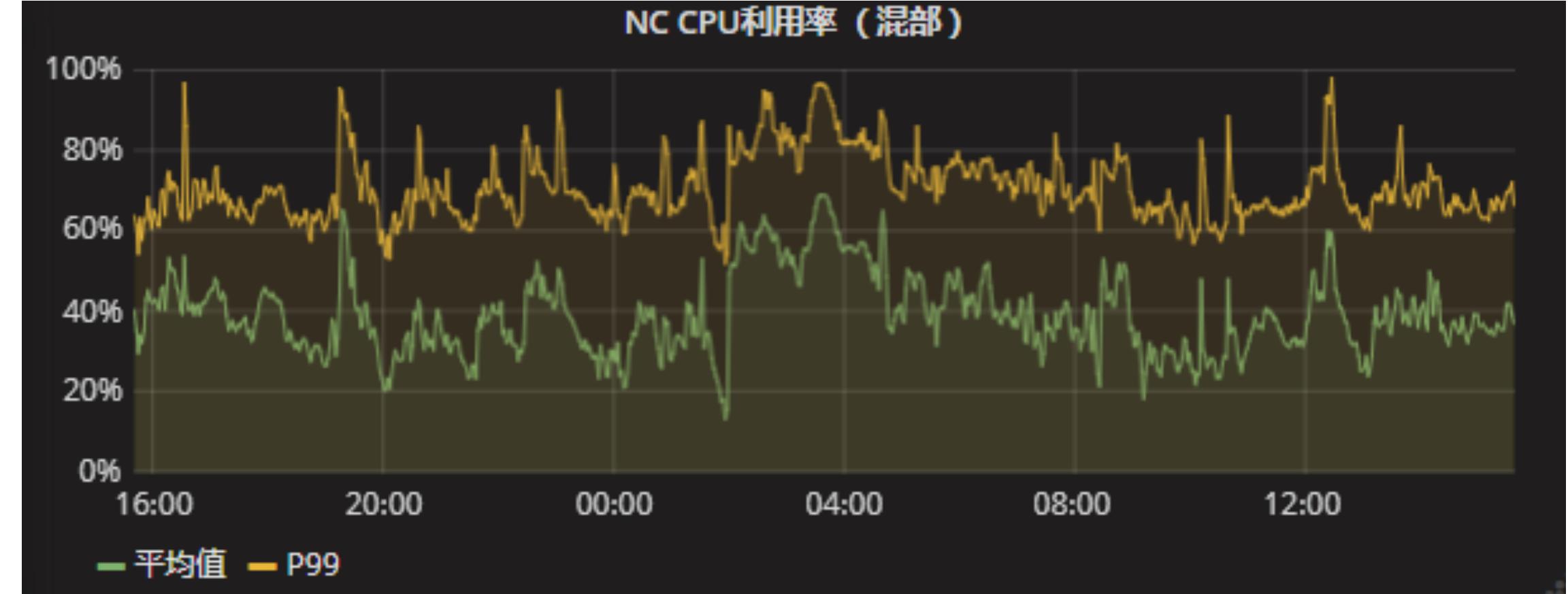
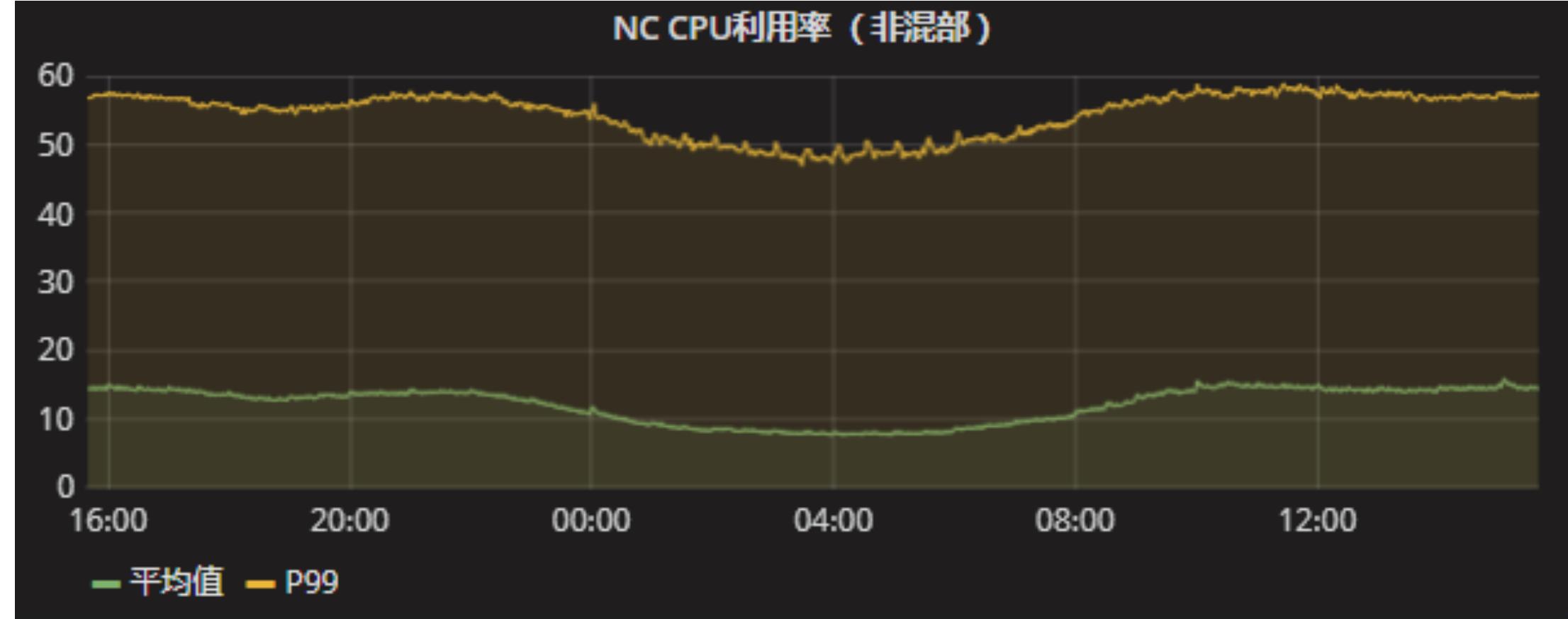
# 混部关键技术

## 在线集群管理

- 应用画像，装箱调度
  - 亲和互斥、任务优先级
  - 稳定性优先、利用率优先
  - 应用自动伸缩、分时复用
  - 整站快速扩缩、弹性内存
- 计算任务调度+ODPS
- 弹性内存分时复用
  - 动态内存超卖
  - 无损降级、有损降级

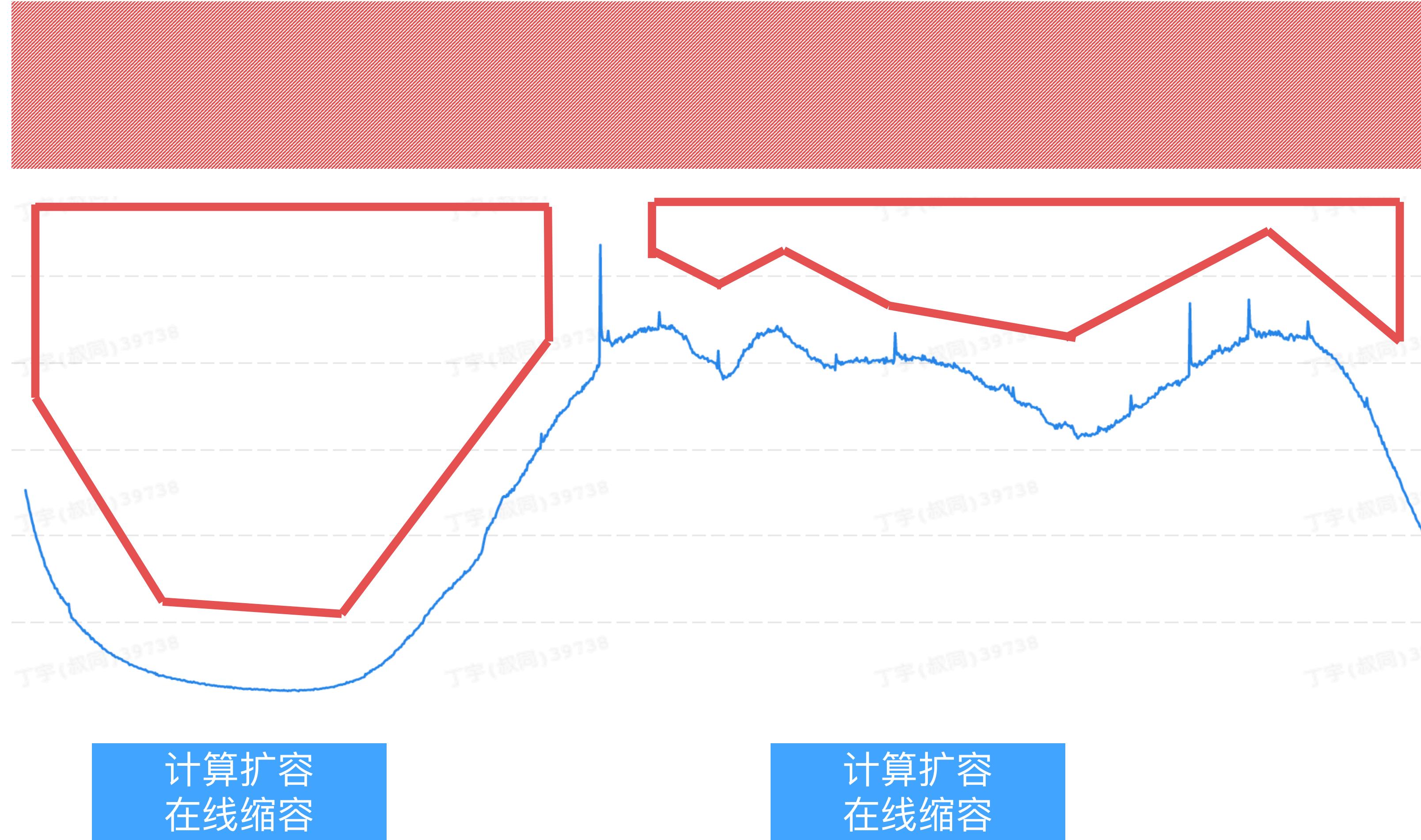


# 混合部署-引入计算任务提升日常资源利用率



- CPU平均利用率10% -> 40%， 延迟敏感类应用RT影响<5%
- 混部集群规模数千台， 经过双11交易核心链路规模化验证
- 为日常节省超过30%的服务器， 明年会扩大10倍部署规模

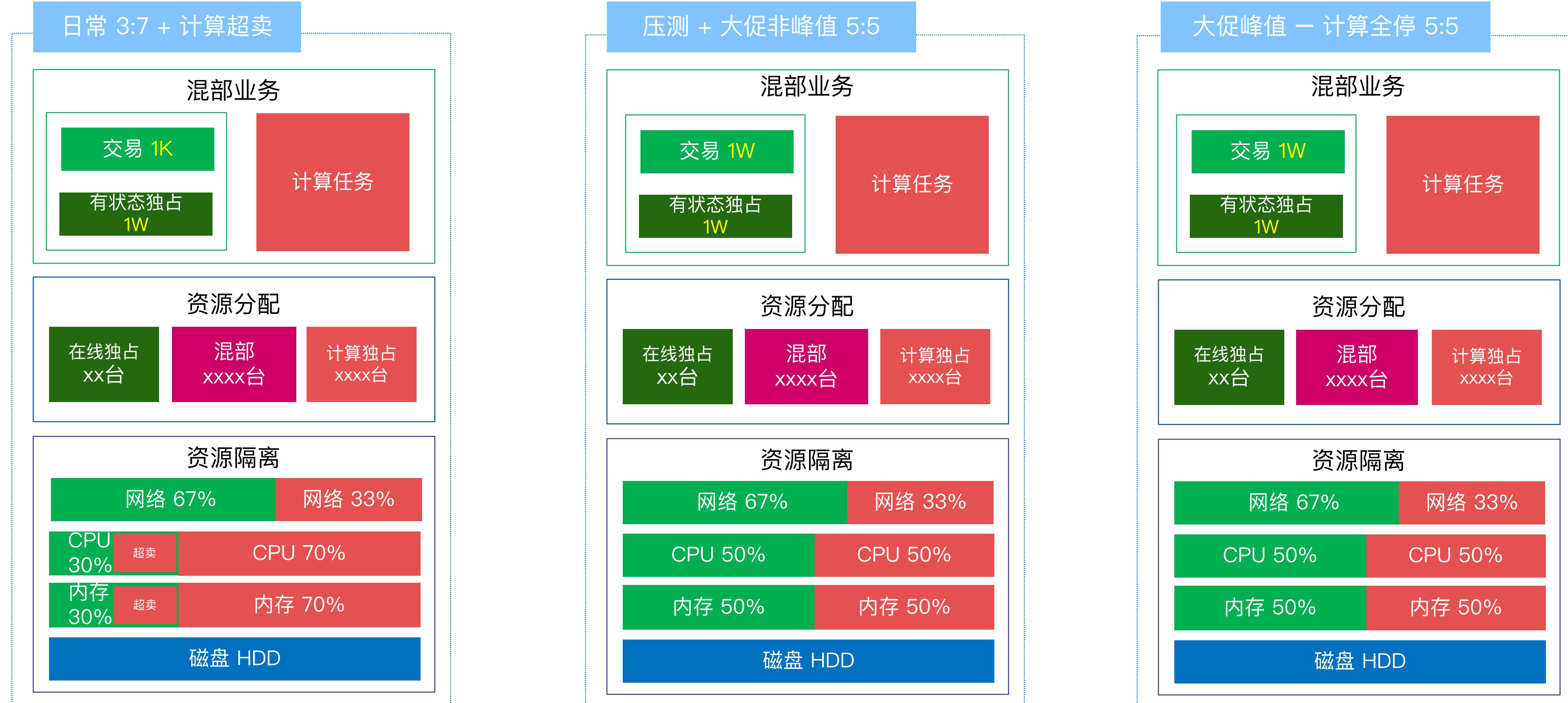
## 混合部署-分时复用进一步提升资源效率



- 时间空间维度优化
- 结合弹性分时复用，  
平均CPU利用率提升  
至60%以上

<https://github.com/alibaba/clusterdata>

# 混合部署-降低大促成本



- 通过部分计算任务短时间降级，空闲资源支持双11交易峰值
- 1小时快速拉起完整站点，大幅降低了双11整体成本

# Pouch简介

- 本意育儿袋，隐喻贴身呵护应用
- 始于2011年，基于LXC，线上大规模应用
- 2015年初开始吸收Docker镜像和标准
- Pouch容器结合AliKernel，大幅增强能力



# Pouch路线选择

- 容器的要素--阿里内部运维和应用视角
  - 有独立IP
  - 能够ssh登陆
  - 独立的文件系统
  - 资源隔离—使用量和可见性



- 手工Hack实现容器要素

- 虚拟网卡, 网桥
- sshd
- Chroot (pivot\_root)
- CGroup, Namespace

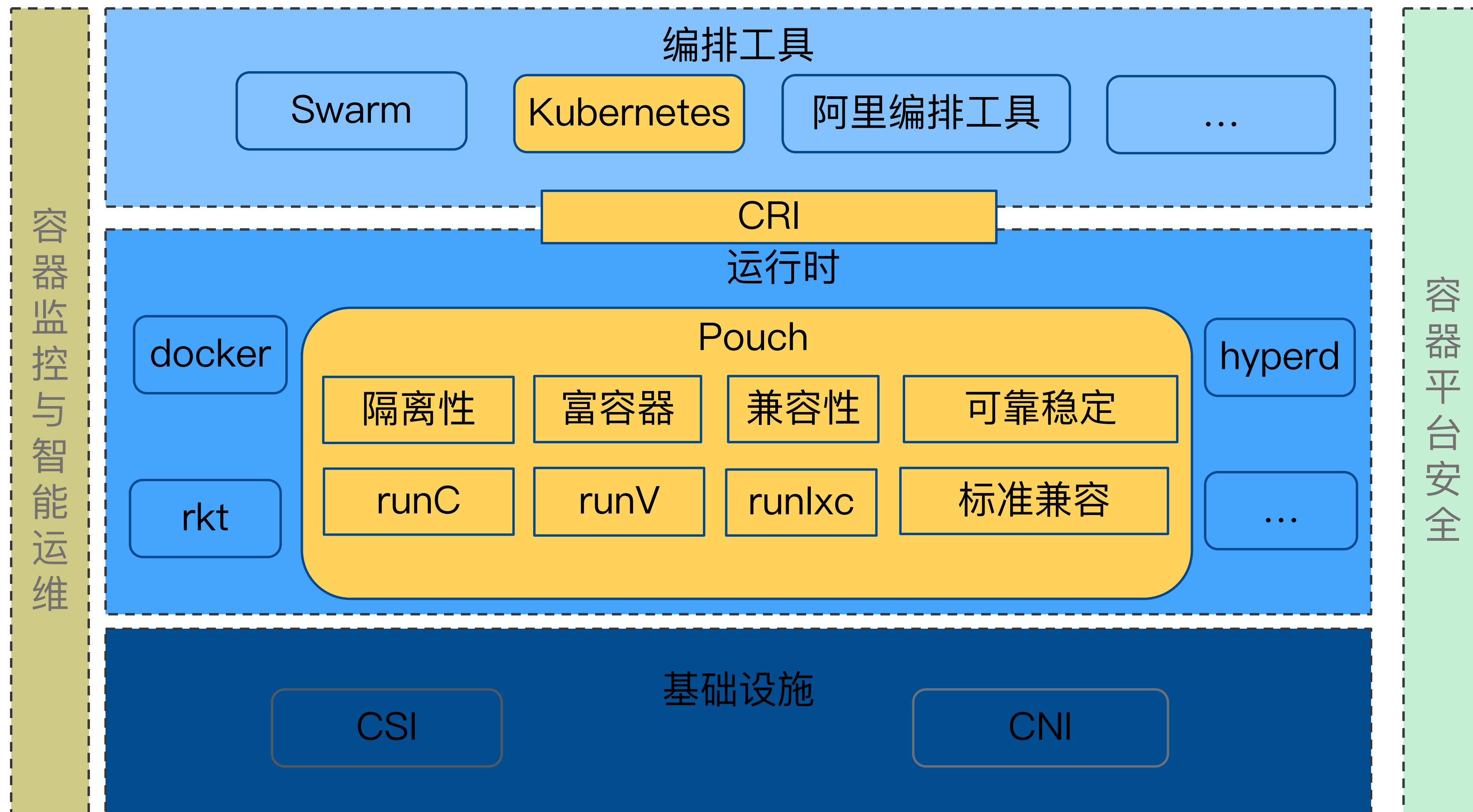


- 引入LXC ([Linux Container](#))
- 内核可见性隔离Patch
- 内核磁盘空间配额Patch

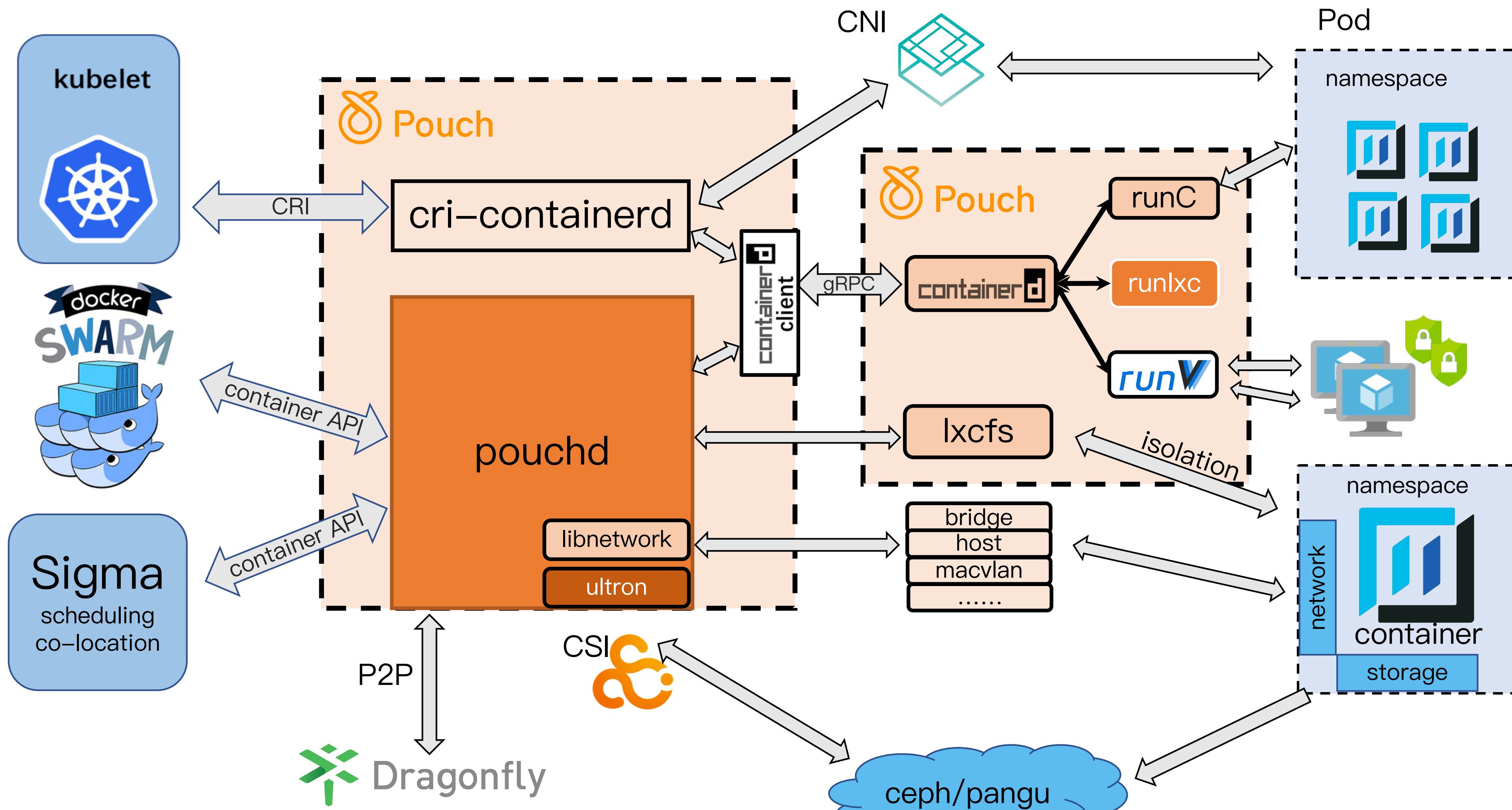


阿里容器技术T4  
引入Docker标准

# Pouch定位



# Pouch架构



# Pouch化进展

## 规模：

- 2017年双11百万级容器
- 在线业务100%容器化
- 计算任务开始容器化
- 拉平异构平台的运维成本

## 覆盖场景：

- 多种编程语言
- DevOps运维体系

## 覆盖业务BU：

- 蚂蚁金服
- 天猫、淘宝
- 合一集团（优酷）
- 菜鸟&高德&UC
- 广告（阿里妈妈）
- 阿里云专有云
- 中间件、数据库

# Pouch开源计划



2017.10.10

合作伙伴共同孵化  
外部开发者邀请内测

2017.11.19

正式开源  
与生态共建Pouch

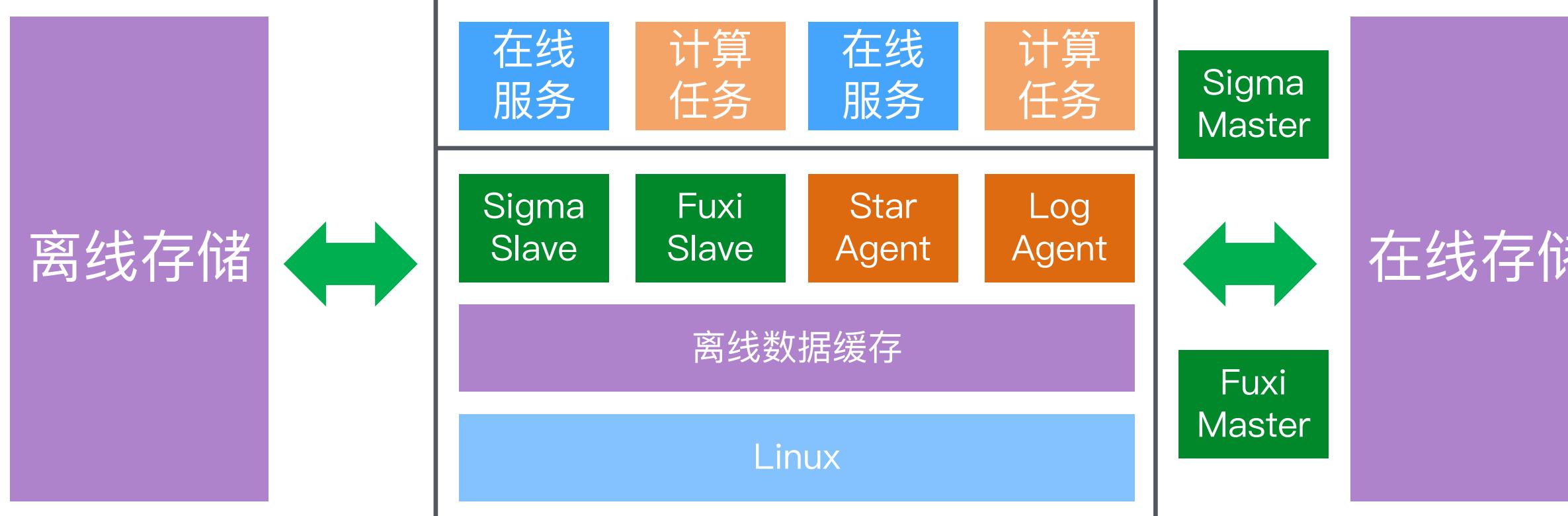
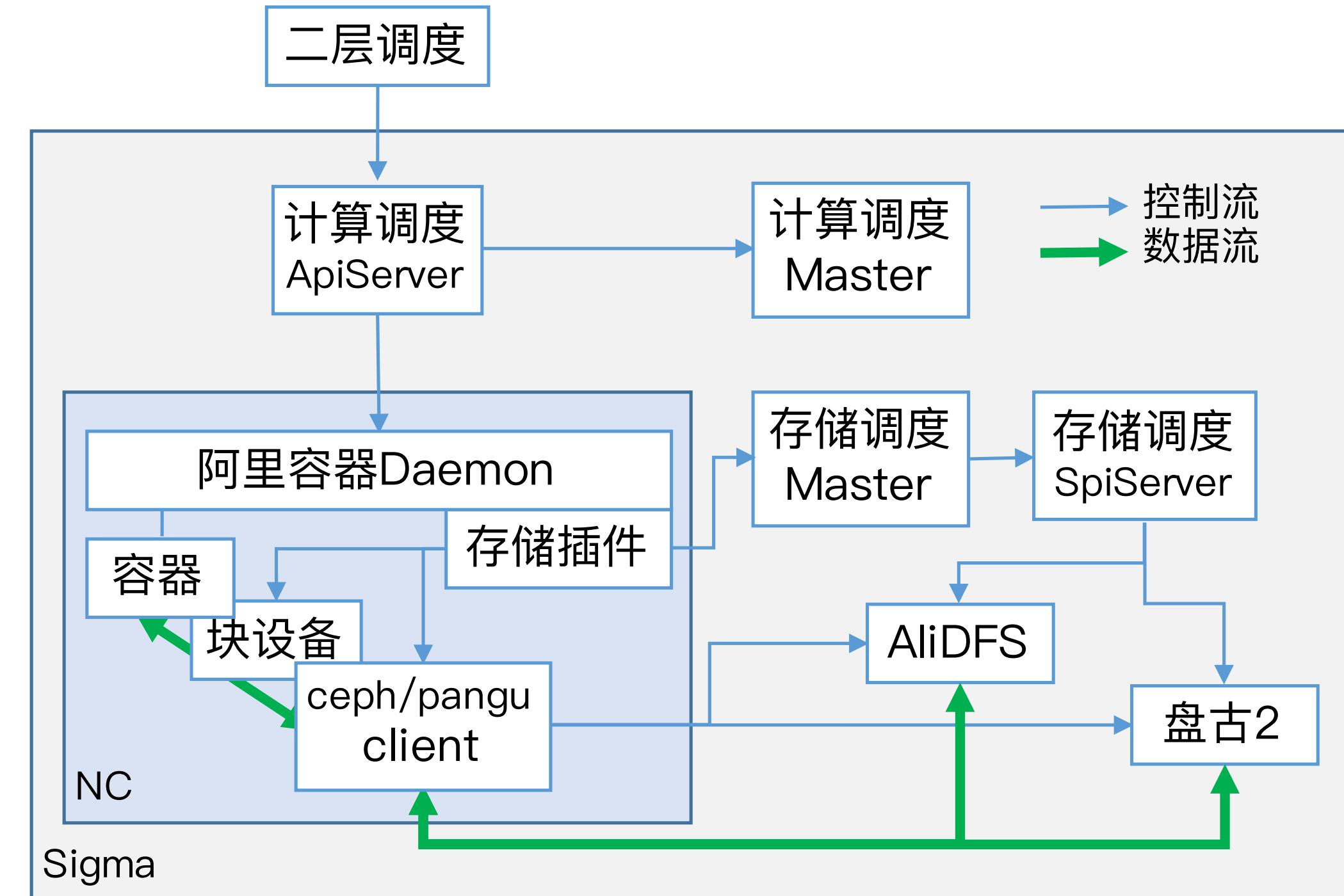
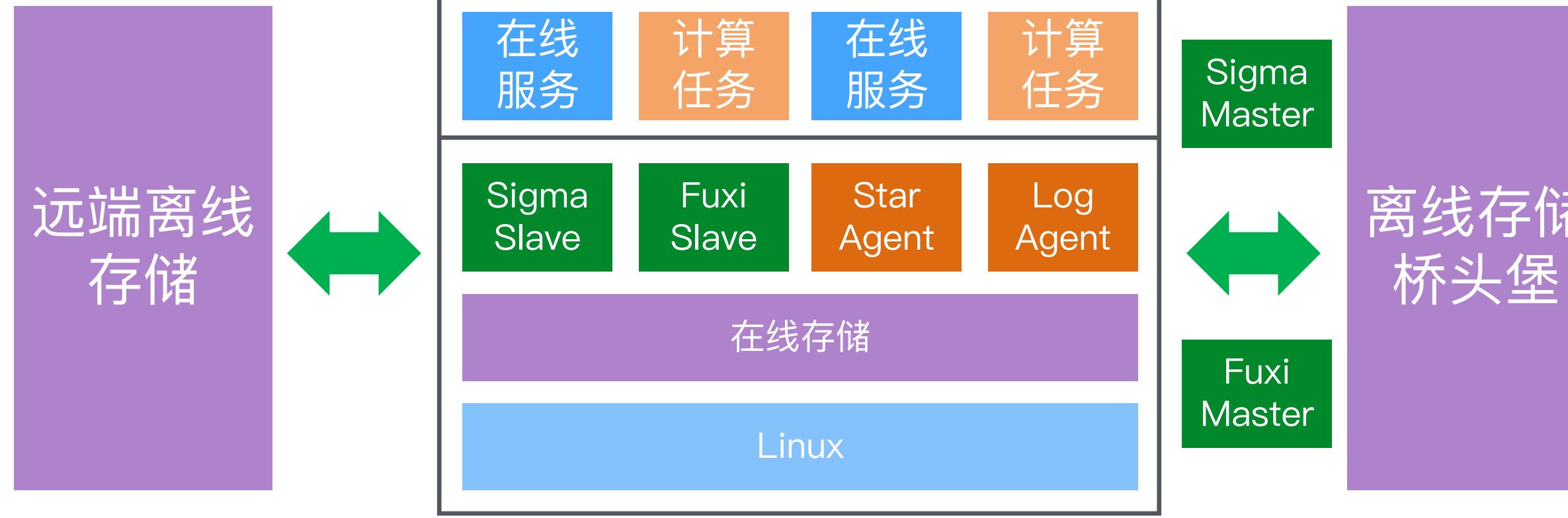
2018.03.31

发布第一个  
大版本

<https://github.com/alibaba/pouch>

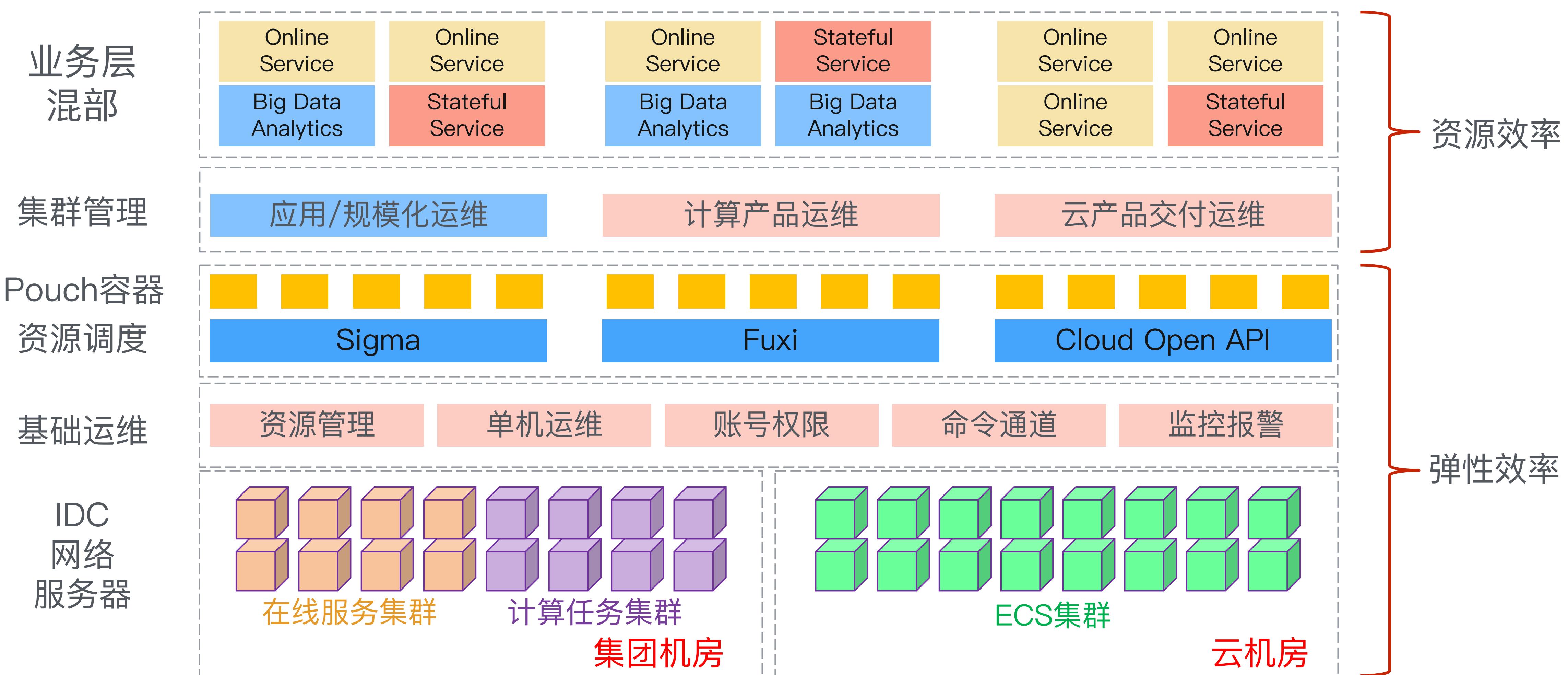
- 推动容器领域发展和标准成熟，成为行业Top3的企业级容器
- 方便传统IT企业利旧，同样享受容器化带来的运维效率优势
- 方便新IT企业享受可靠稳定和多标准兼容的优势

# 存储计算分离



- 不受网络长传带宽限制
- 大集群减少跨网络核心对穿流量
- 网络架构升级、25G、overlay

# 云化架构技术体系

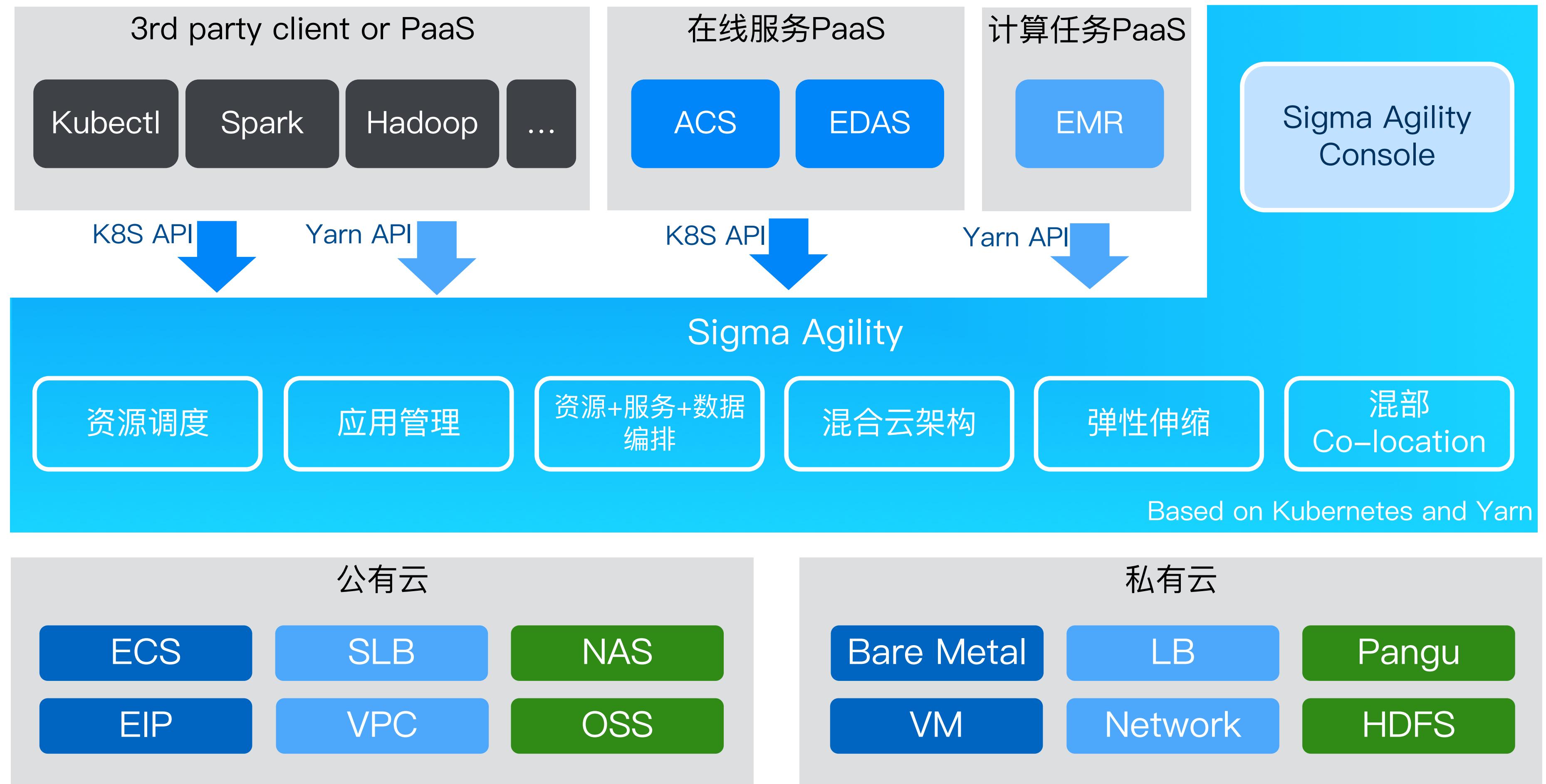


- datacenter as a computer, 多个数据中心像一台计算机一样来管理, 可以跨多个不同的平台来调度业务发展所需的资源
- 构建混合云以极低成本拿到服务器, 解决有没有的问题, 通过弹性分时复用和混部大幅提升资源利用率, 解决好不好的问题
- 真正实现弹性资源平滑复用、任务灵活混合部署, 用最少服务器、最短时间、最有效率完成容量目标
- 通过云化架构使双11新增IT成本下降50%, 使日常IT成本下降30%, 带来容器、调度和集群管理领域的技术价值爆发

# 调度体系技术方向

- 最优利用：最大化资源利用率
  - 统一调度，全面容器化，计算存储分离，大DC：奠定规模效应基础
  - 混合部署：不同资源使用特性的业务共享资源
  - 分时复用：集群弹性，不同峰值时间分布的业务共享资源
  - 公共云弹性：效率大幅优化，解决大促、日常突增和脉冲需求
  - 精细化调度：更好的了解应用，增强隔离，提高利用率分配率、降低碎片率
- 最优规划：资源运营优化
  - 统一资源池、合并Buffer，统一资源管理
  - 打通资源需求、规划和分配，全局整体优化，减小库存
  - 提升在线率、统一预算、统一交付、交付分时复用、过保延期下线

# 调度体系产品化



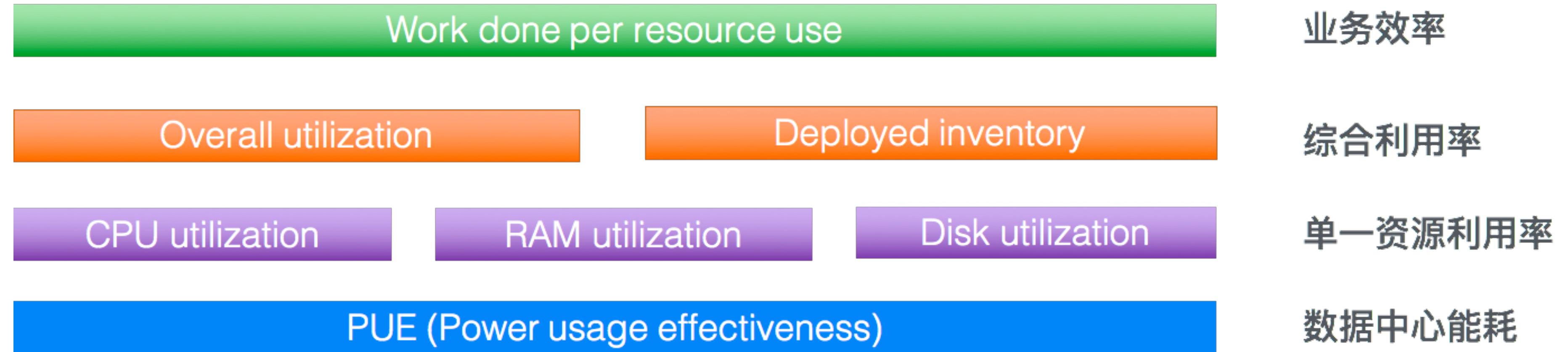
## 定位

- 兼容Kubernetes架构和标准
- 阿里内部调度、容器、运维领域优势技术产品化
- 提供企业级容器应用管理能力，提高企业IT效率

## 优势

- 混部 (Co-location)
- 混合云资源管理和建站
- 灵活的调度策略和算法
- 与阿里云生态无缝集成
- 经过双11大规模场景检验

# IDC利用率



# Thanks

---

系统软件事业部 打造具备全球竞争力、效率最优的系统软件