ArchSummit
全球架构师峰会 2017

主办方 Geekbang 极客邦科技 InfoQ

阿里巴巴云化架构创新之路

丁宇（叔同）

SPEAKER INTRODUCE

丁宇（叔同）

2017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

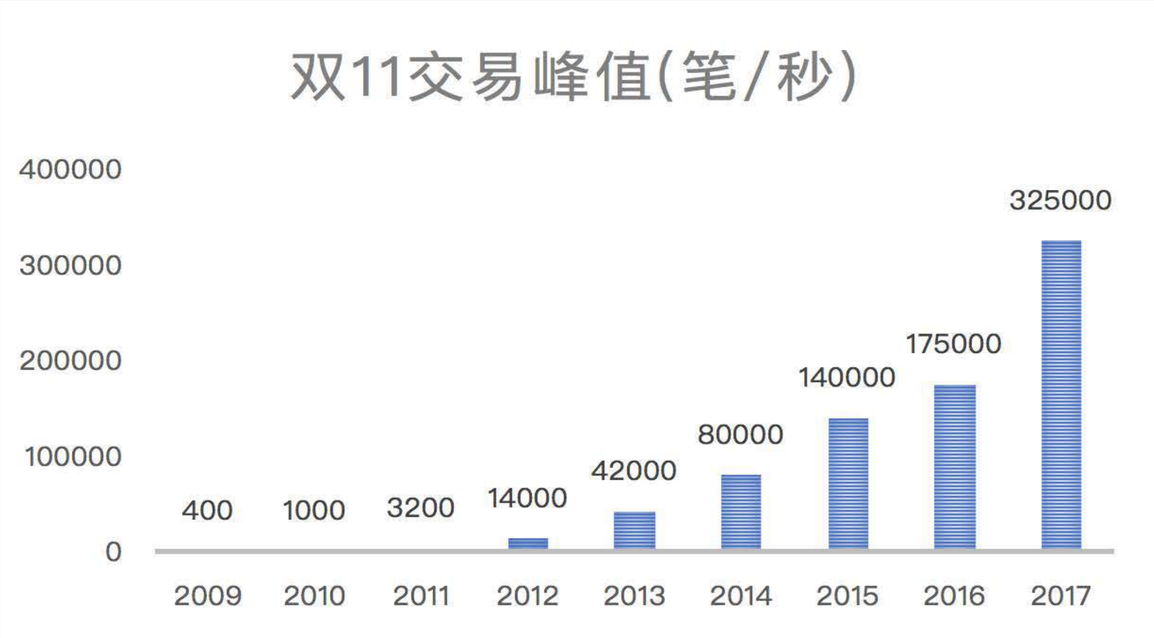
联系方式：18657182390

TABLE OF

CONTENTS 大纲

- 双11的技术挑战与突破
- 云化架构演进的背景
- 统一调度和混部的挑战
- Pouch容器和容器化的进展
- 云化架构和双11的未来技术路线

双11的技术挑战



- 双11的技术挑战，互联网级的规模，企业级的复杂度，金融级的稳定性，数十倍的业务峰值
- 9次双11交易额增长280倍，交易峰值增长800多倍，系统复杂度和大促支撑难度以指数级攀升
- 双11峰值的本质是用有限成本去最大化的提升用户体验和整体吞吐能力，用合理的代价解决峰值
- 发挥规模效应，持续降低单笔交易成本以提升峰值能力，为用户提供丝般顺滑的浏览和购物体验

双11的技术突破

扩展性问题

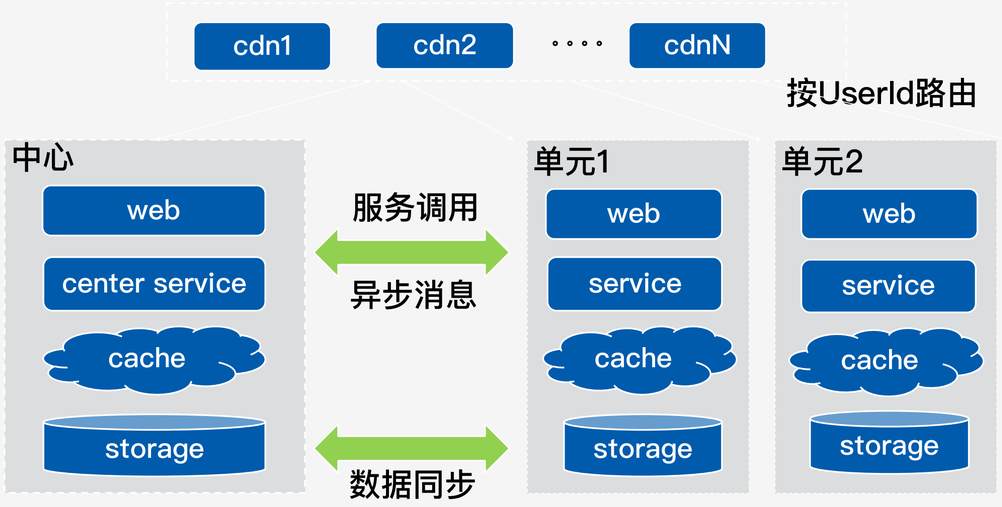
- 分布式架构
- 异地多活

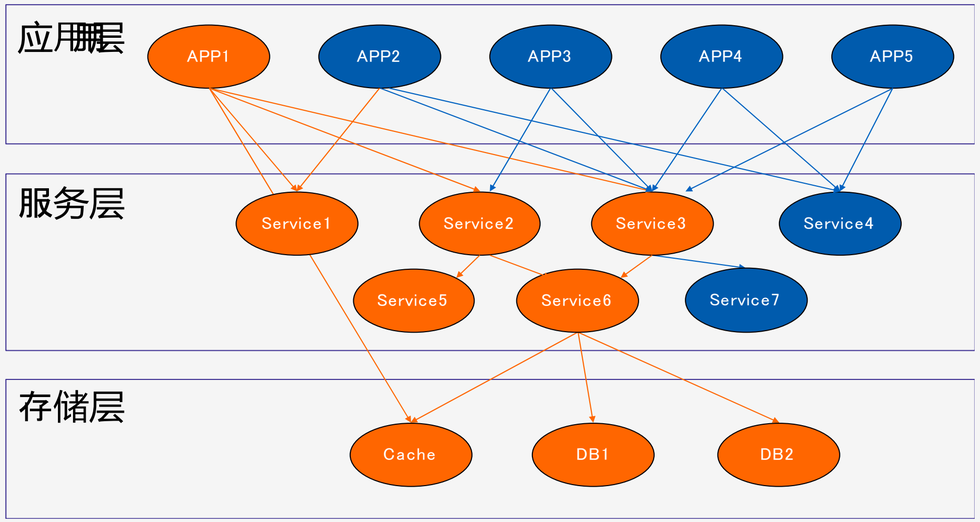
稳定性问题

- 限流降级
- 全链路压测

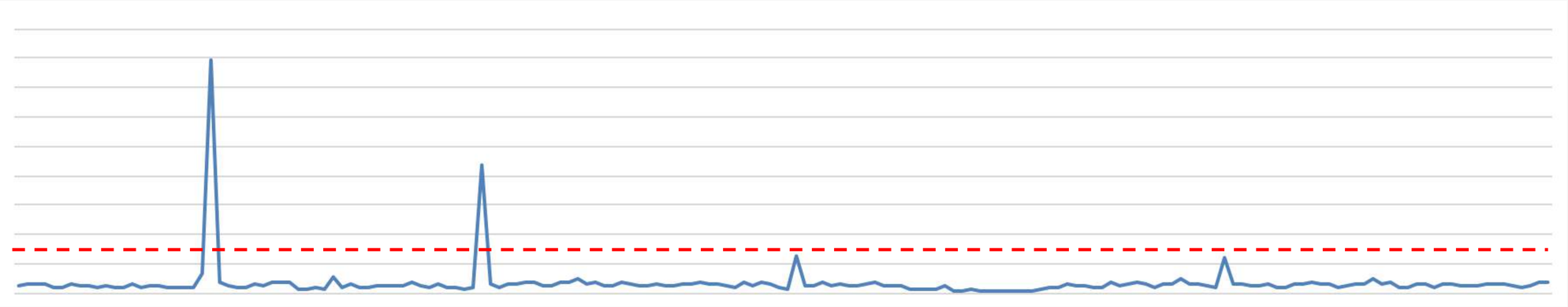
新的技术挑战

- 成本、效率



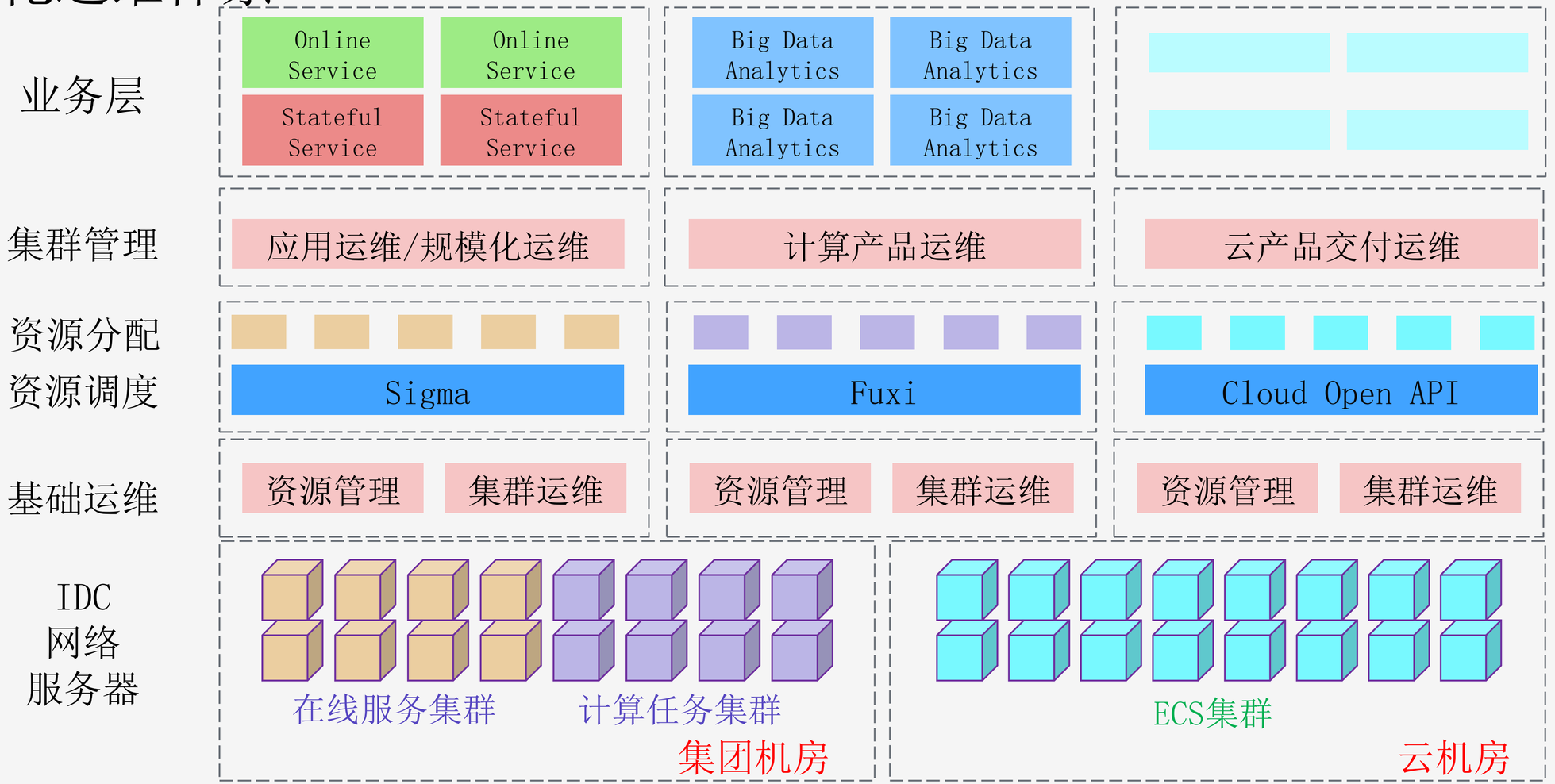


云化架构演进背景



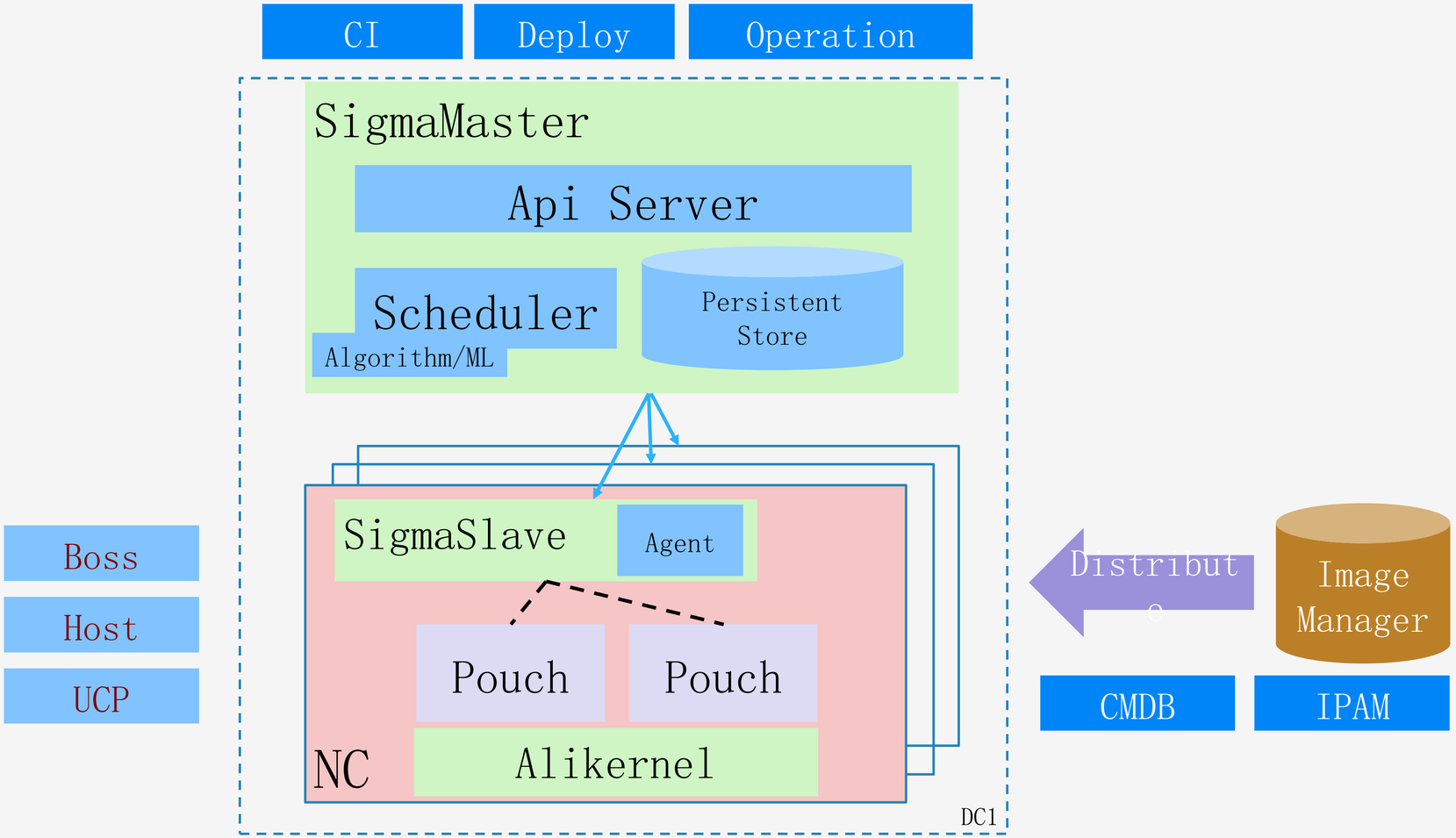
- 双11只有一天，过后资源利用率不高，隔年会形成较长时间的低效运行
- 资源整体弹性能力不足，运维体系差异大，各版块无法平滑复用
- 每个版块有不同的Buffer池，在线率、分配率、利用率无法统一
- 通过云化架构提升整体技术效率，提高全局资源弹性复用能力
- 拉通技术体系，降低大促和日常整体成本，双11单笔交易成本减半

垂直化运维体系



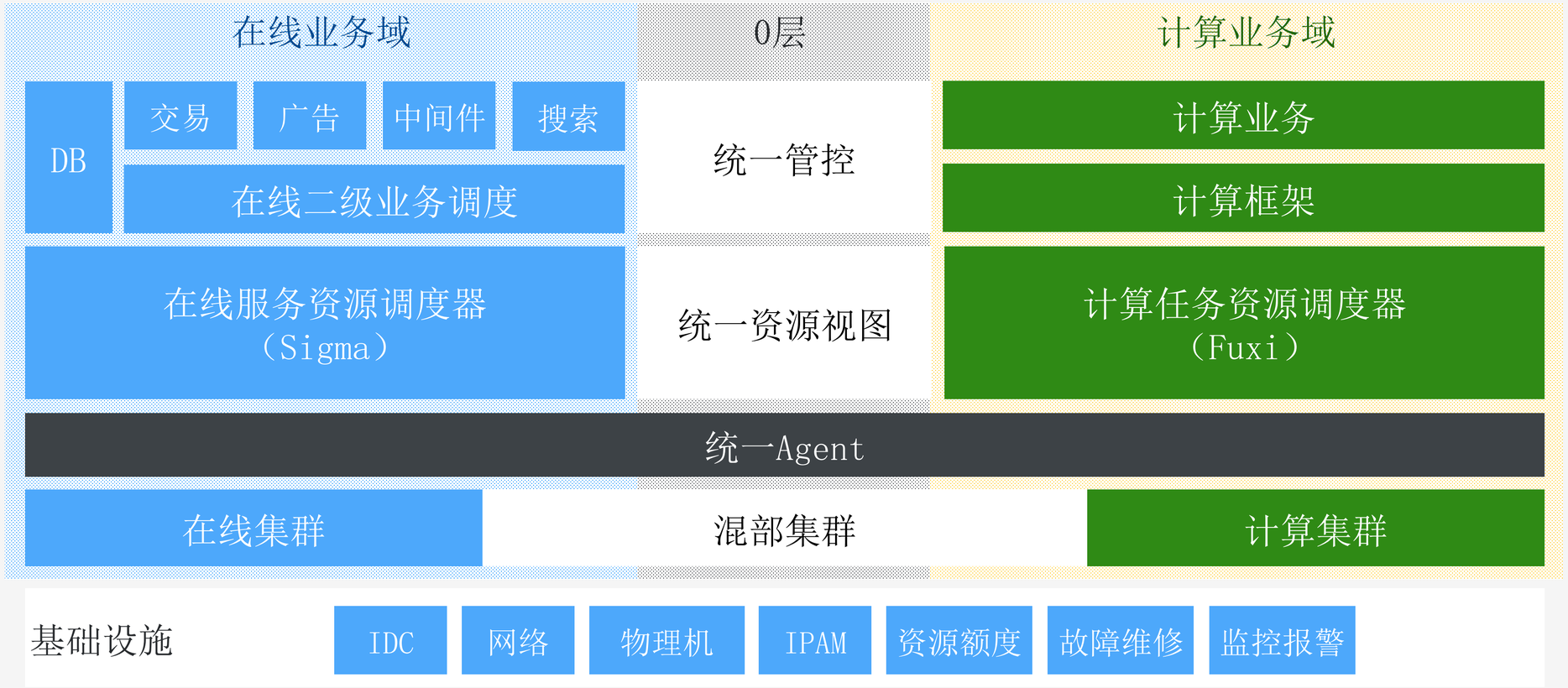
- 技术全面云化，逐层重构升级，弹性复用资源，全局统一调度，在线服务和计算任务混部
- 统一运维部署、资源分配的标准，提高调度效率，容量自动交付，全面容器化
- 充分发挥云计算的弹性能力，减少自采基础设施投入，建设混合云，一键建站

集群管理和调度系统Sigma



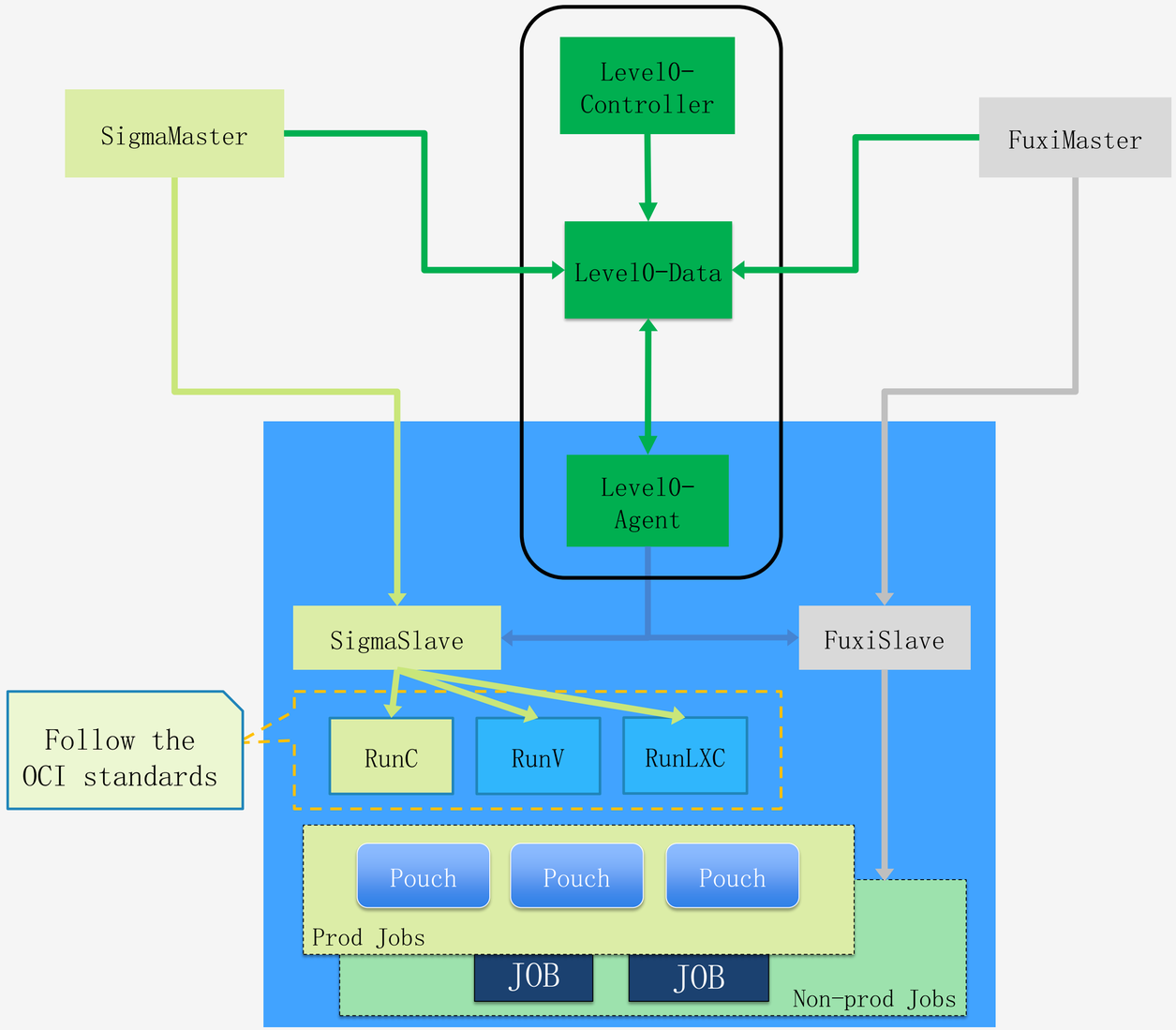
- 始于2011年，以调度为中心的集群管理体系
- 面向终态的架构设计；三层大脑合作联动管理
- Go语言重构，17年兼容Kubernetes API，和开源社区共同发展

调度优化



- 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

混部关键技术

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

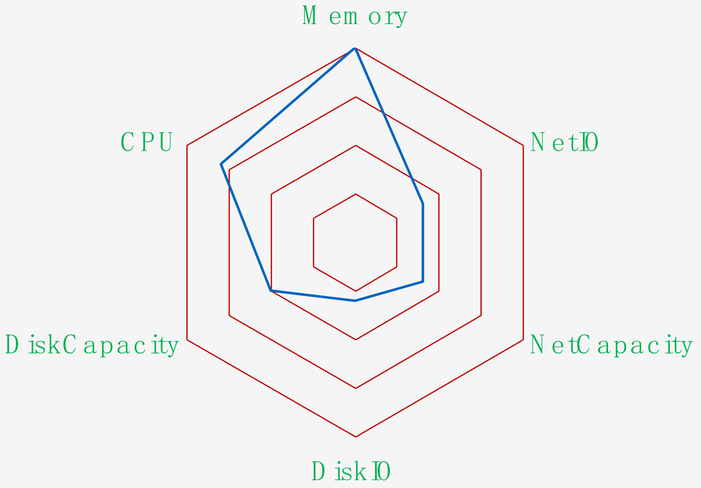
混部关键技术

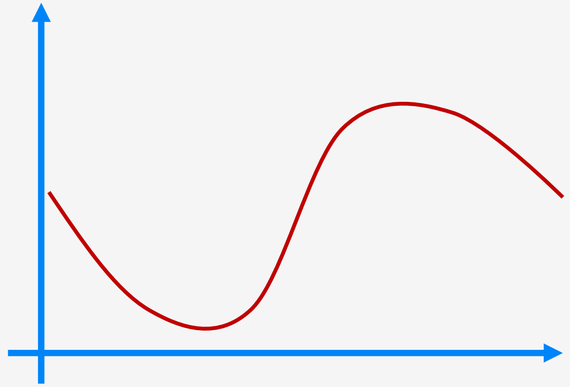
在线集群管理

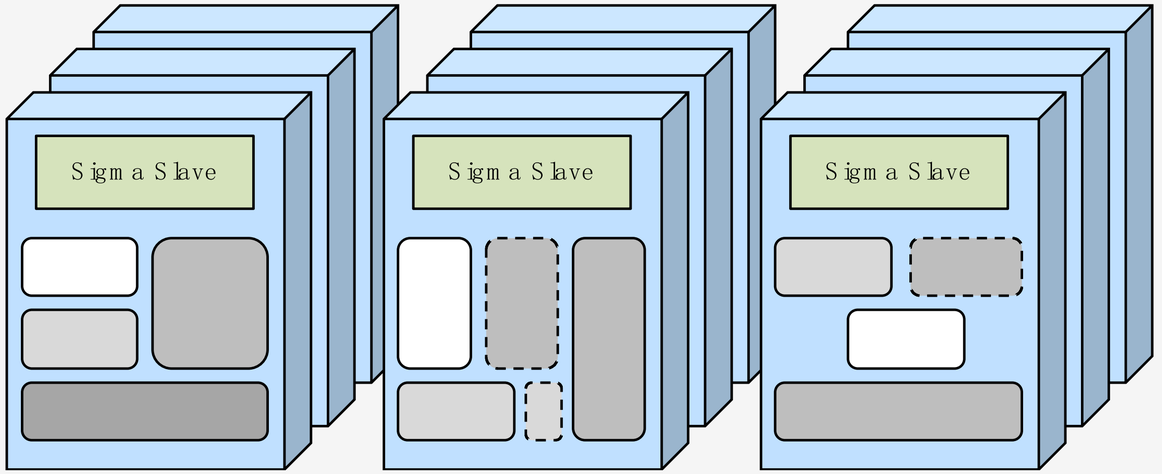
- 应用画像，装箱调度
- 亲和互斥、任务优先级
- 稳定性优先、利用率优先
- 应用自动伸缩、分时复用
- 整站快速扩缩、弹性内存

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效率

- 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

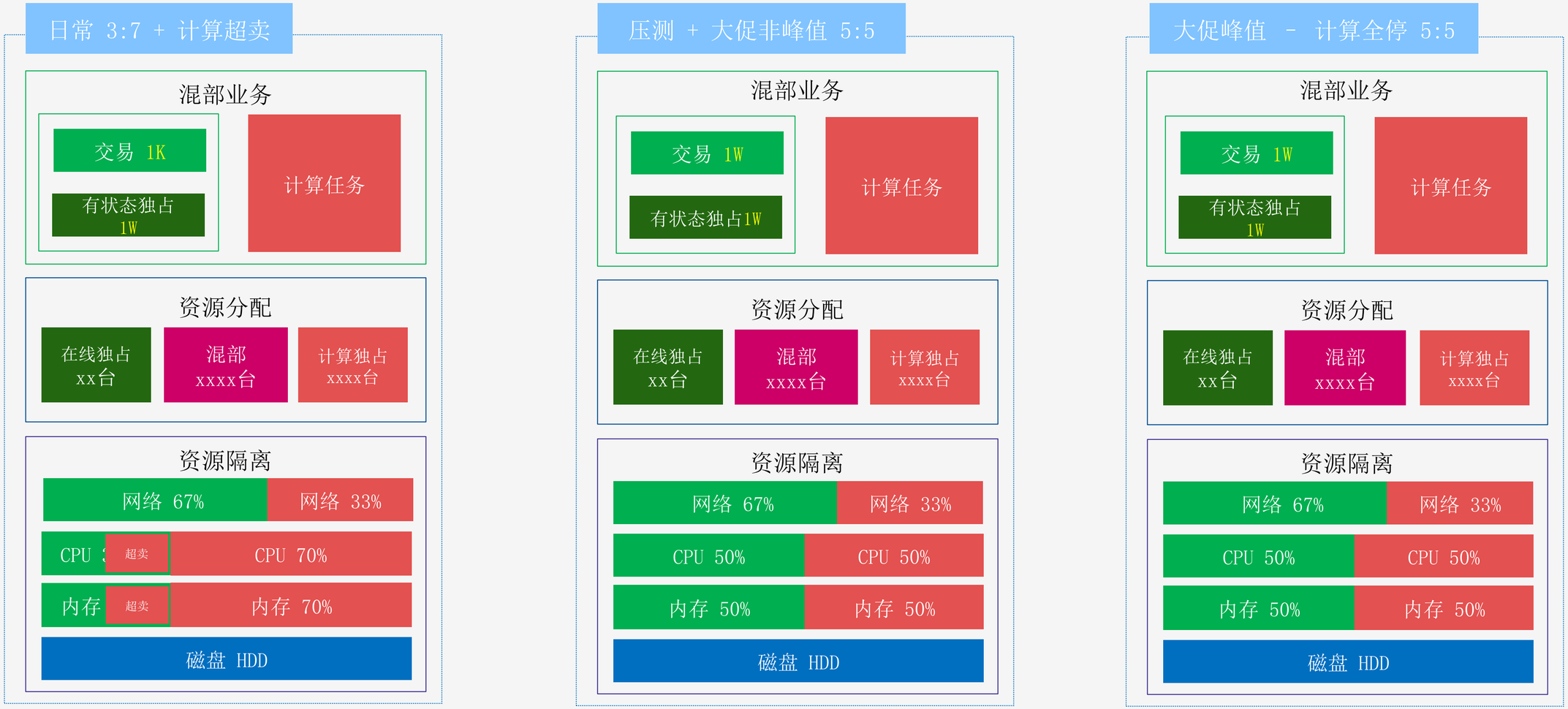
计算扩容
在线缩容

计算扩容
在线缩容

- 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发展路线

- 容器的要素--内部应用运维视角
  - 有独立IP
  - 能够ssh登陆
  - 独立的的文件系统
  - 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

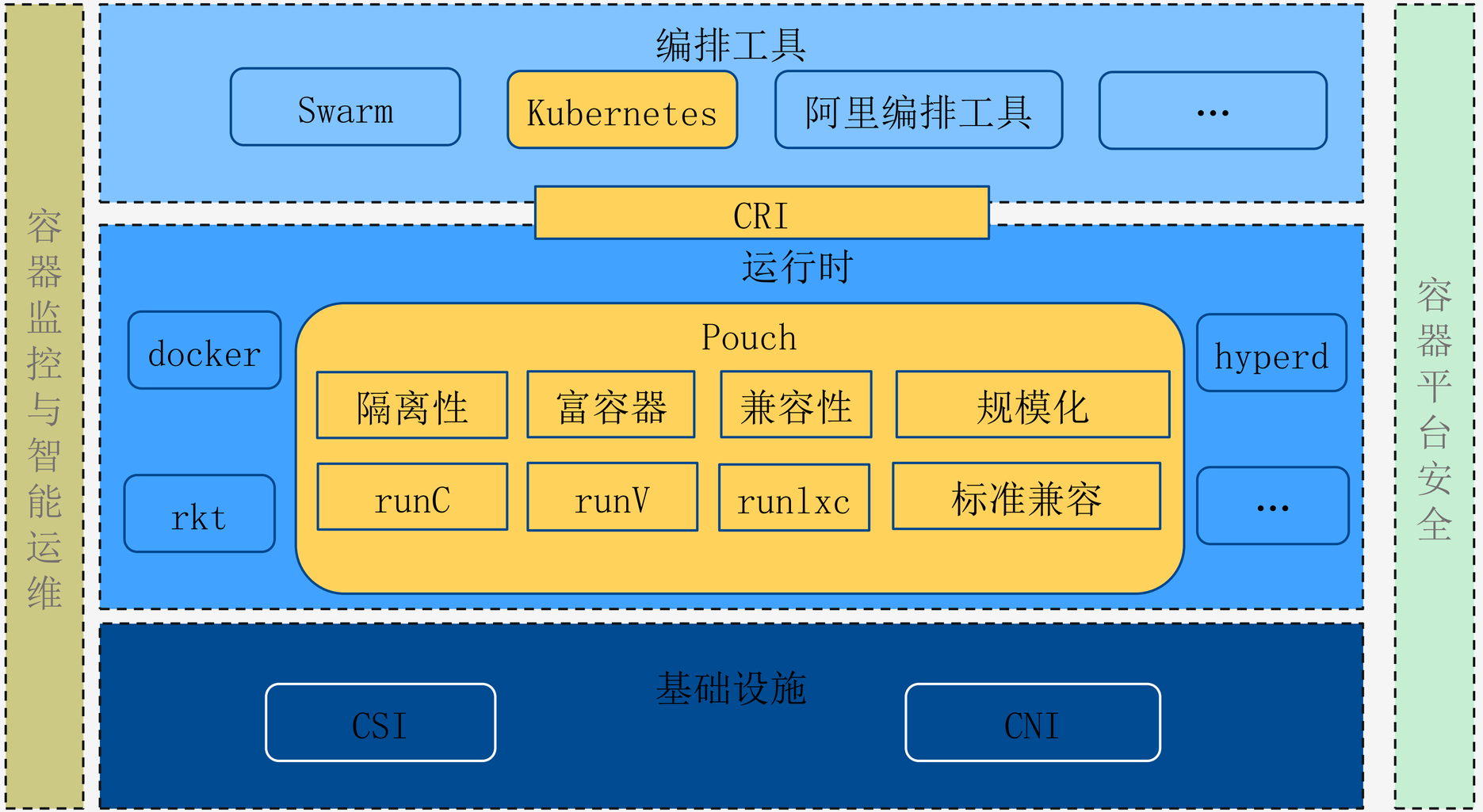
- 引入LXC（Linux Container）
- 内核可见性隔离Patch
- 内核磁盘空间配额Patch

阿里容器技术T4

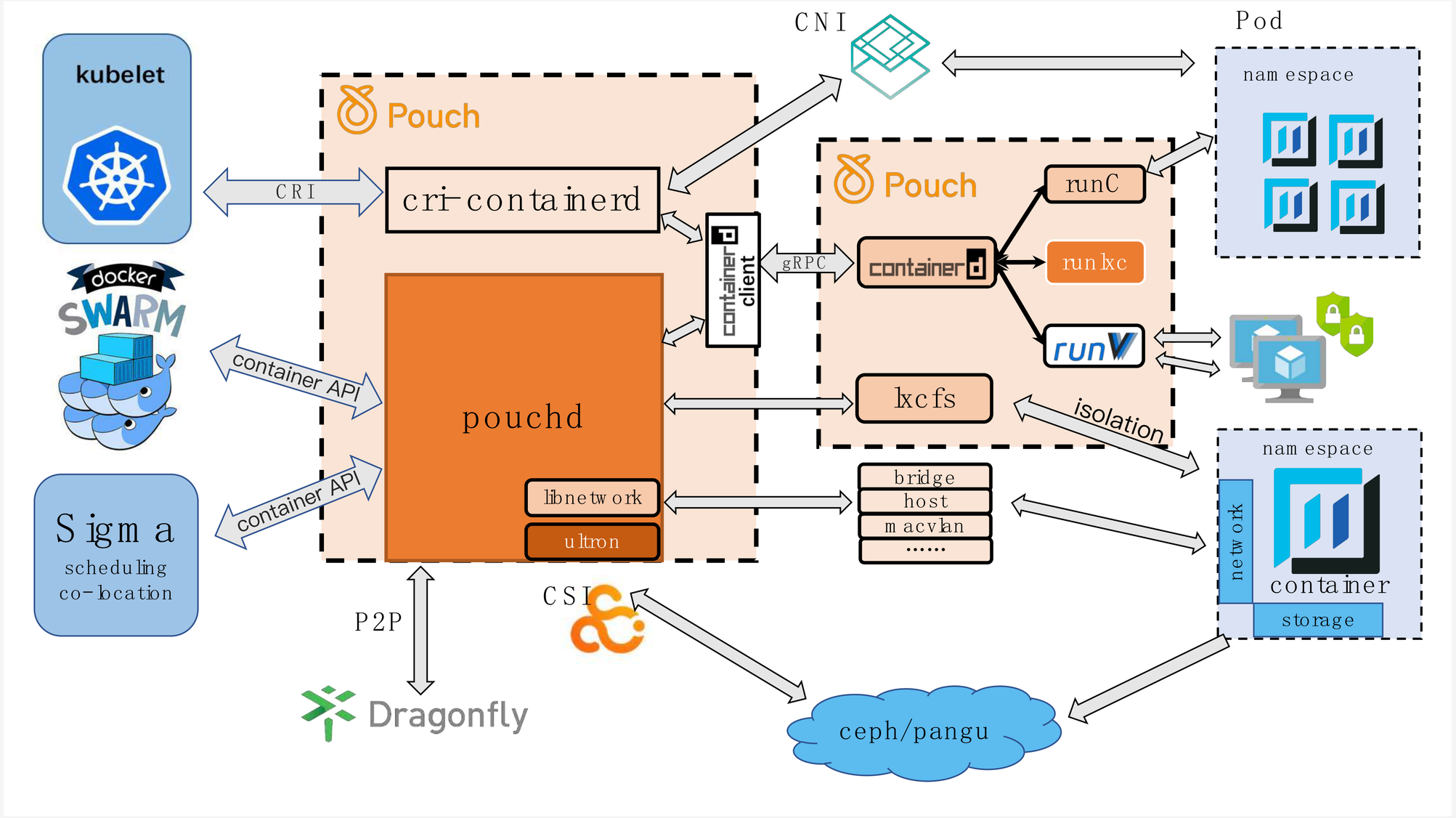
引入Docker标准

Pouch

Pouch定位


容器监控与智能运维
编排工具
Swarm
Kubernetes
阿里编排工具
...
CRI
运行时
docker
rkt
Pouch
隔离性
富容器
兼容性
规模化
runC
runV
runlxc
标准兼容
hyperd
...
基础设施
CSI
CNI
容器平台安全

Pouch架构


kubelet
docker SWARM
Sigma
scheduling
co-location
CRI
container API
container API
P2P
Dragonfly
Pouch
cri-containerd
pouchd
libnetwork
ultron
container client
gRPC
CNI
CSI
Pouch
containerd
runC
runlxc
runV
lxcfs
isolation
bridge
host
macvlan
.....
Pod
namespace
namespace
network
container
storage
ceph/pangu

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
宣布开源
外部开发者内测孵化

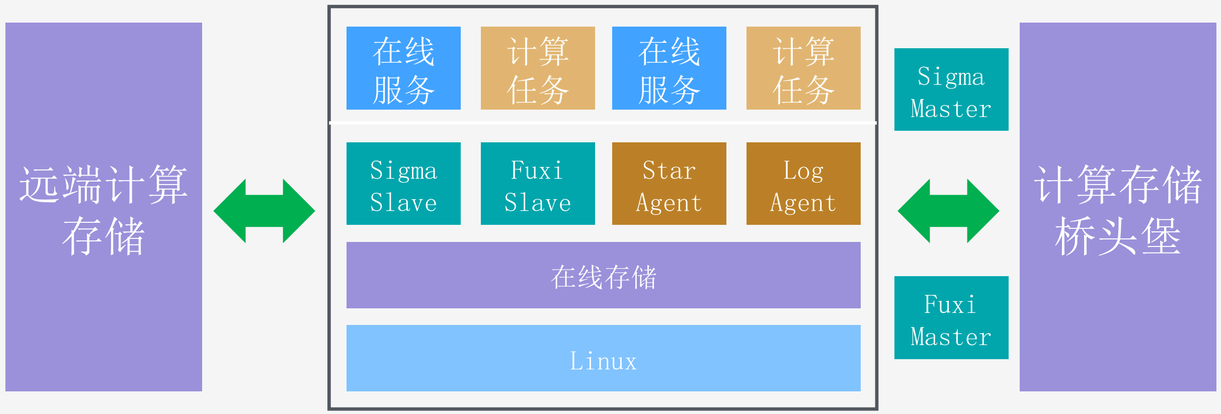
2017.11.19
正式开源
与生态共建

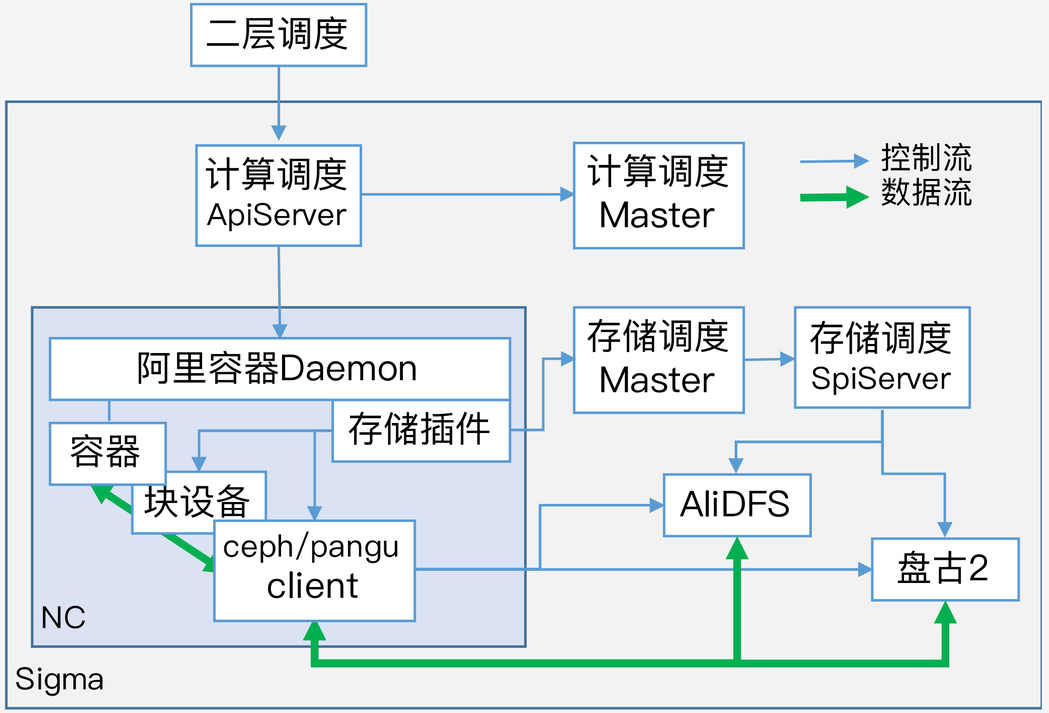
2018.03.01
发布第一个
大版本

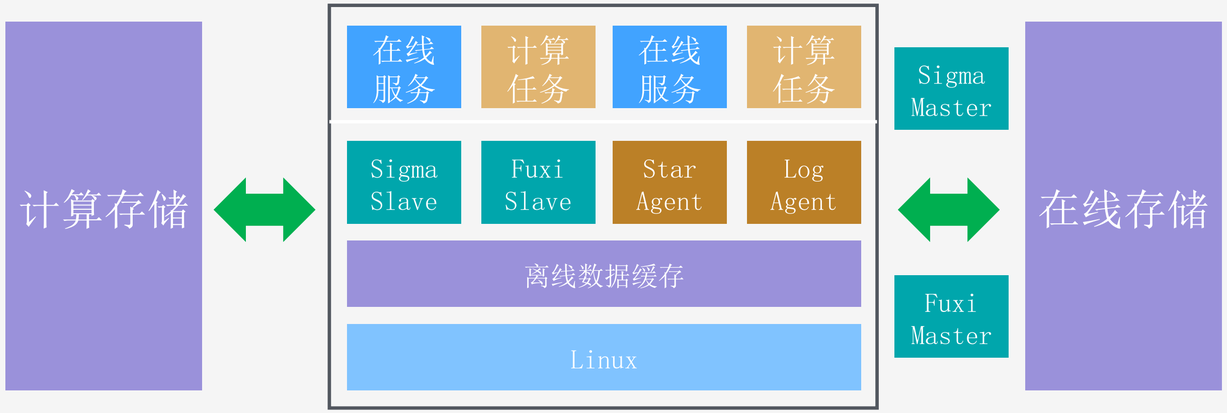
https://github.com/alibaba/pouch

- 推动容器领域发展和标准成熟，给业界提供差异化有竞争力的选择
- 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企业享受规模化、稳定性和多标准兼容的优势

存储计算分离

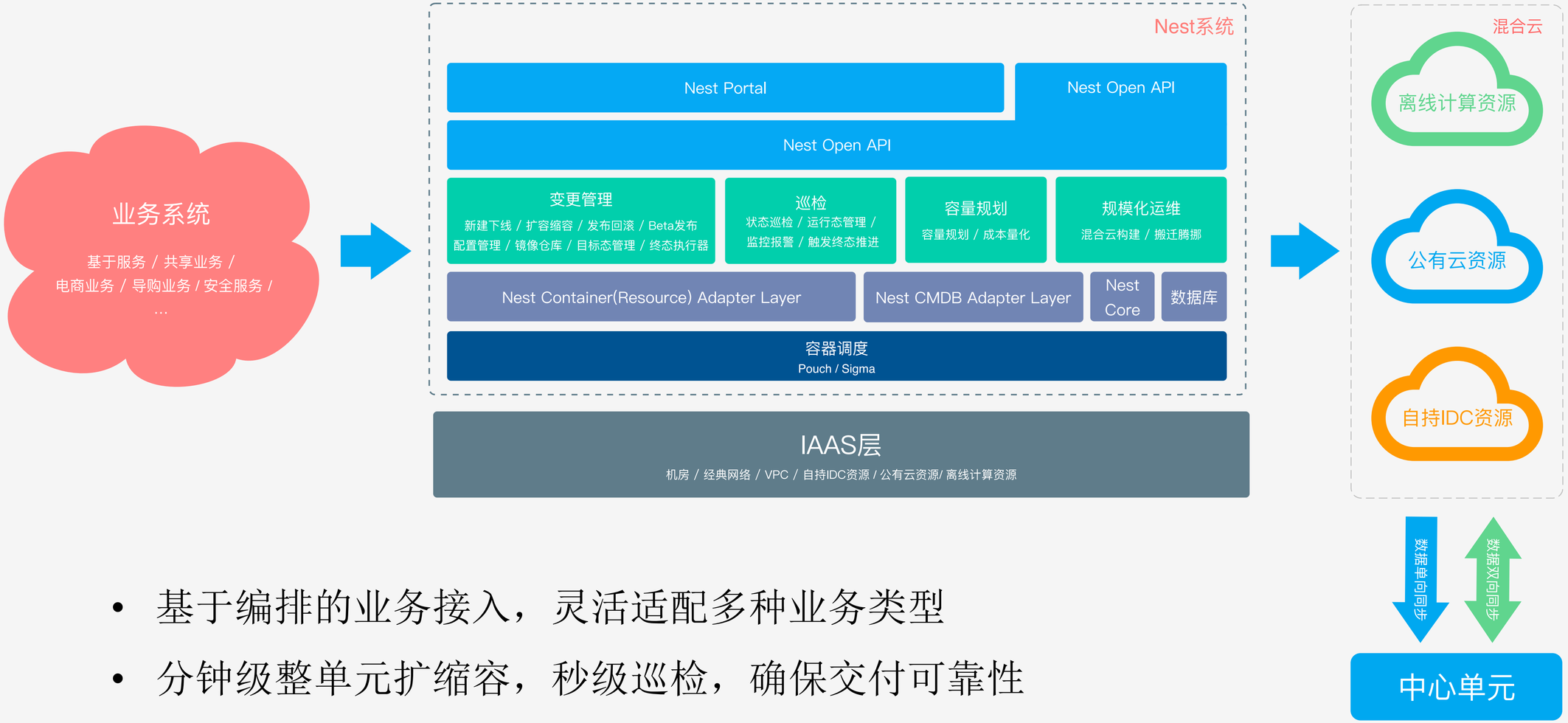






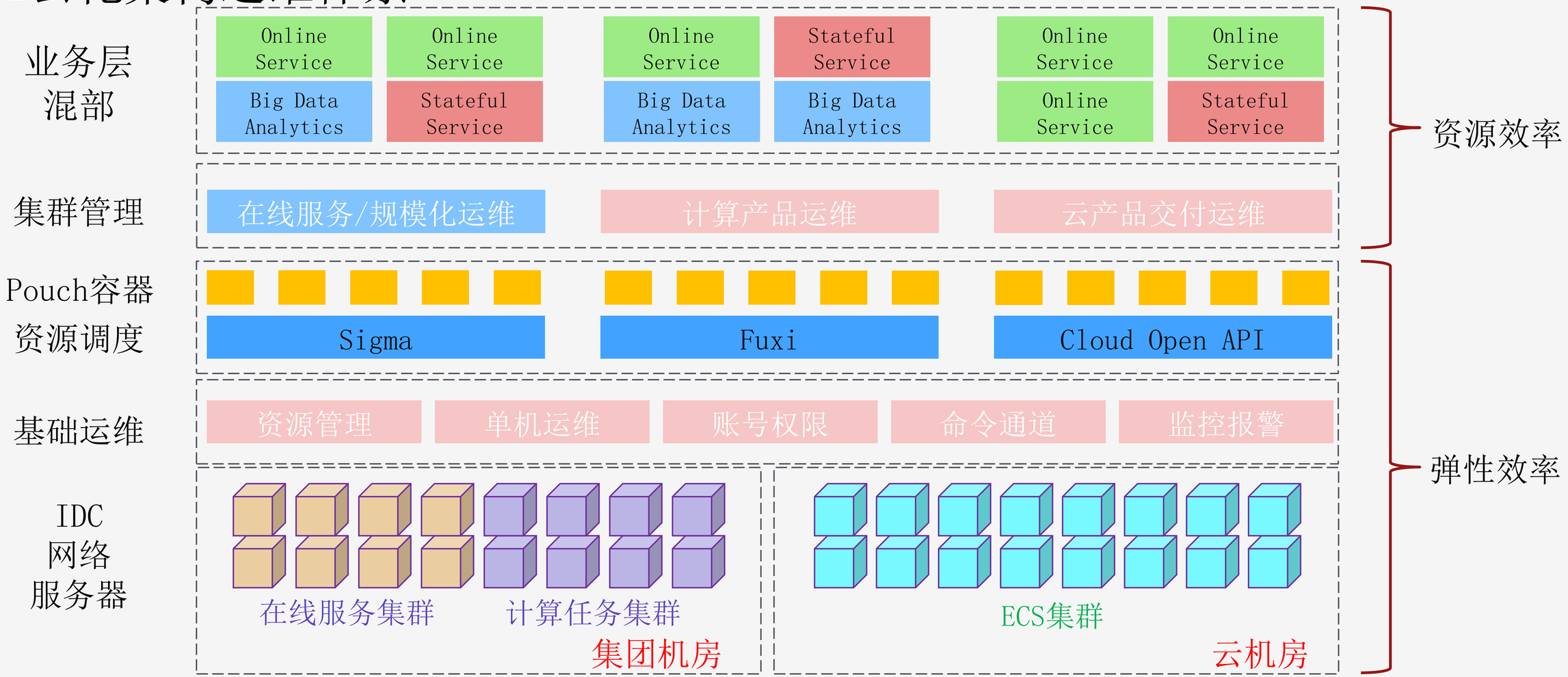
- 不受网络长传带宽限制
- 大集群减少跨网络核心对穿流量
- 有状态服务的存储计算分离
- 网络架构升级、25G、overlay

混合云弹性架构



- 基于编排的业务接入，灵活适配多种业务类型
- 分钟级整单元扩缩容，秒级巡检，确保交付可靠性
- 降低资源持有时间和非online时间，提升弹性效率
- 双11全面使用阿里云基础设施，8小时快速构建全球最大混合云

双11云化架构运维体系



- 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、最优效率完成容量目标
- 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

Sigma Agility

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

云化架构及双11未来的思考

- 提升IDC资源利用率，扩大弹性规模扩充混部形态，继续释放标准化规模化的技术红利
- 面向终态的体系结构升级，产品化输出赋能行业，持续提升IT效率、降低社会创新成本

- 更低的成本来实现系统的可扩展性，自动化的方式来确保正确性跟稳定性
- 加速基础技术迭代，效率、成本、体验和最大吞吐能力找到新的平衡点
- 数据算法驱动、智能决策处理，提升整个系统的运营效率及用户体验
- 提升双11准备和作战的效率，人与机器智能协同指挥

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全球架构师峰会 2017