GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

会议时间：3月25日-3月26日

会议地点：深圳 • 南山区 圣淘沙酒店(翡翠店)

主办单位：开放运维联盟 OOPSA Open OPS Alliance　高效运维社区 GreatOPS Community

指导单位：数据中心联盟 Data Center Alliance

协办单位：中国新一代IT产业推进联盟

GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

京东Docker实践

何小峰 京东/云平台

议题

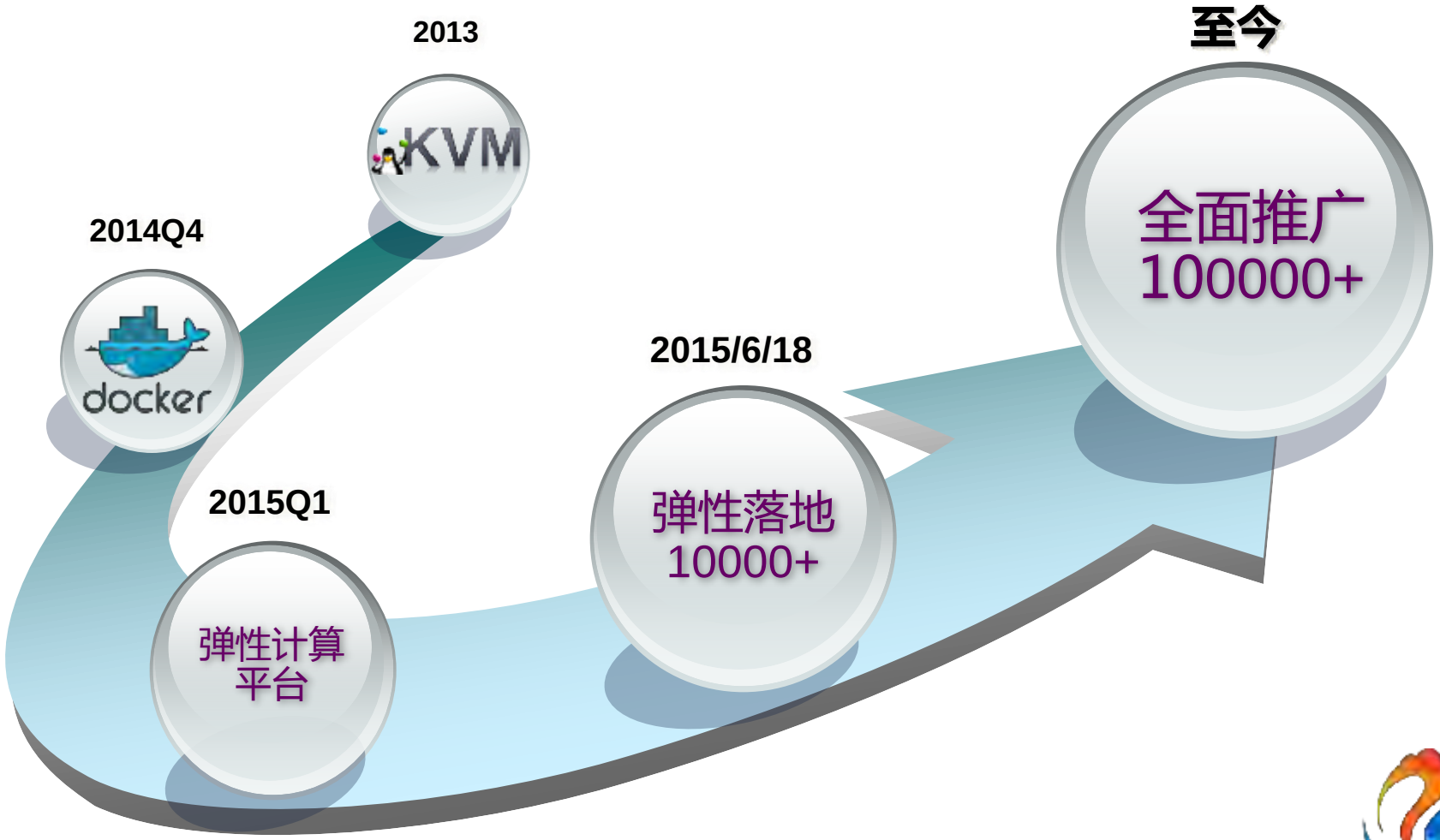
1. 京东容器之路
2. 弹性计算架构
3. 弹性计算应用场景
4. 自动化运维
5. 数据驱动的精细化运营

面临的挑战

- 硬件采购周期长，交付效率不高；
- 不能准确评估资源使用情况，无法精细化运营；
- 硬件成倍增长，成本高；
- 扩容慢，压力来的时候不能快速扩容；
- 部署环境复杂，运维压力大；

用户关注

容器化之路

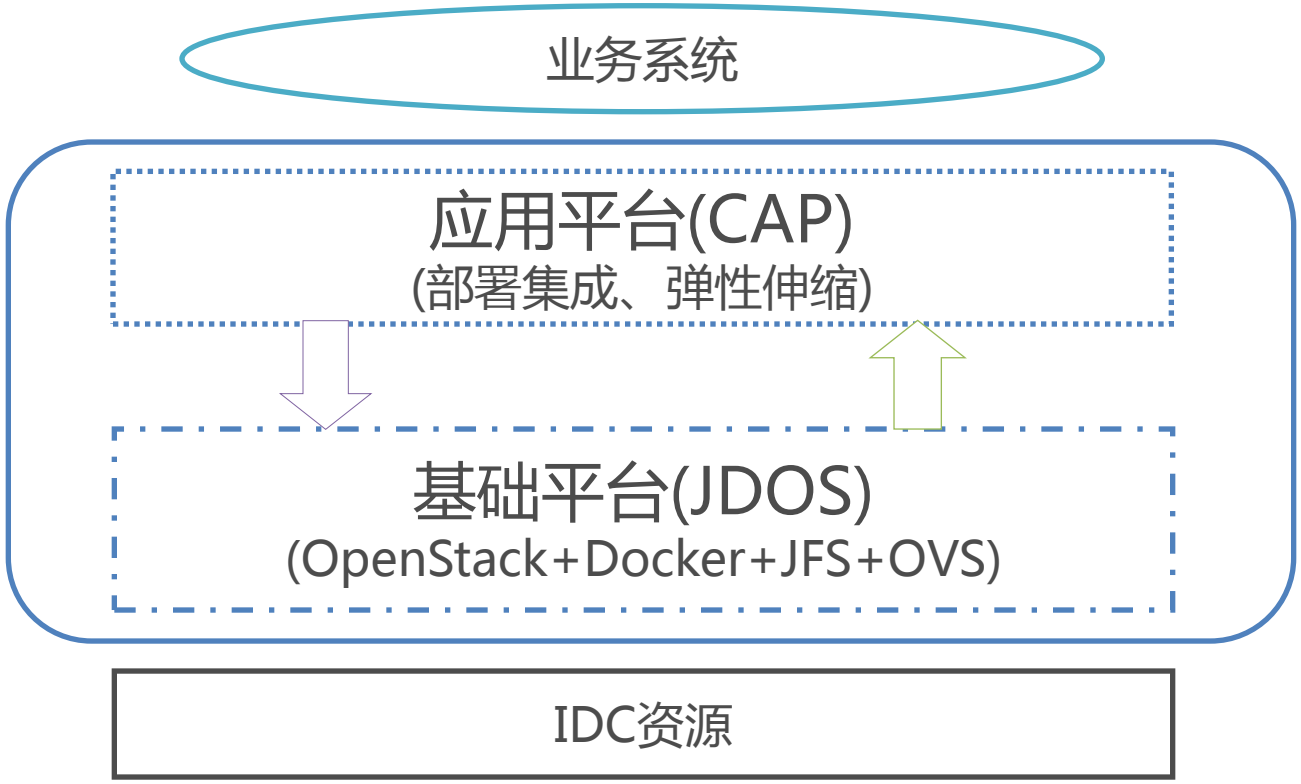

2013
KVM
至今
全面推广
100000+
2014Q4
docker
2015/6/18
弹性落地
10000+
2015Q1
弹性计算
平台

选择Docker的原因

议题

1. 京东容器之路
2. 弹性计算架构
3. 弹性计算应用场景
4. 自动化运维
5. 数据驱动的精细化运营

弹性计算架构

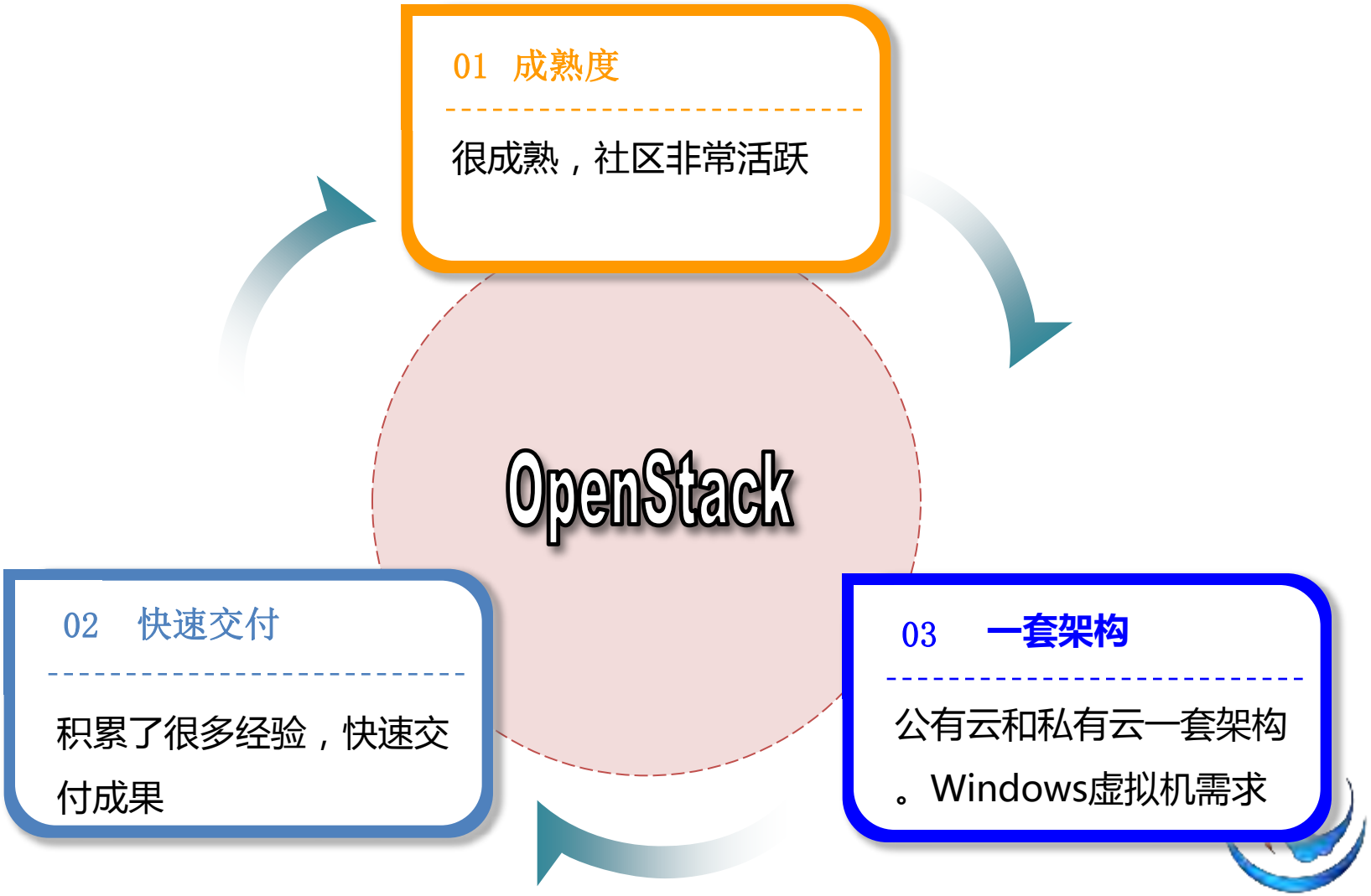


弹性计算平台 = JDOS（JD Datacenter OS）+CAP（Cloud Application Platform）。

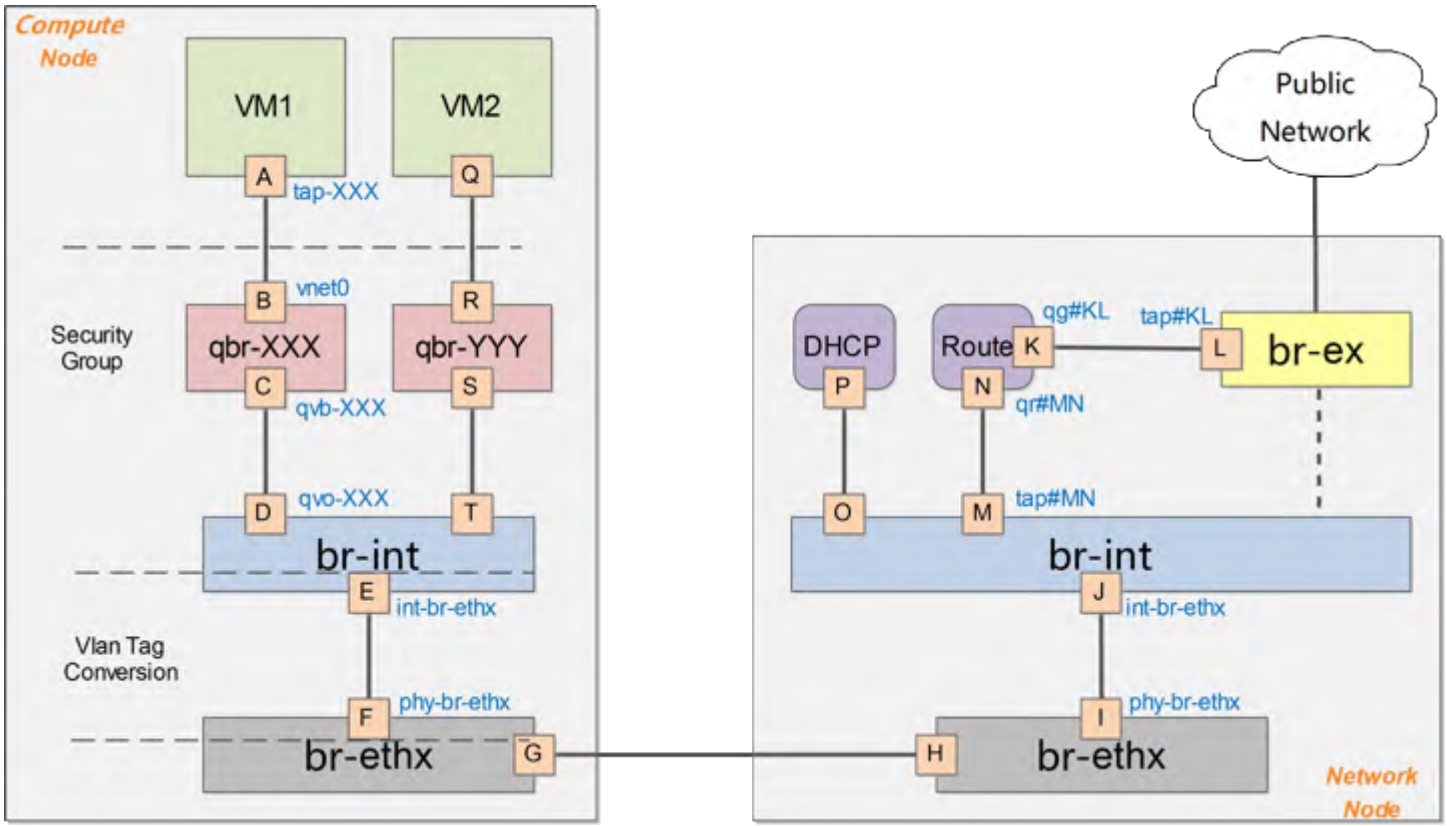
■JDOS实现实现基础设施（网络，物理机，存储）的资源管理、容器的生命周期管理、监控指标采集；

■CAP负责应用治理、部署、监控报警、资源利用率统计、手动和自动的弹性伸缩。

OpenStack

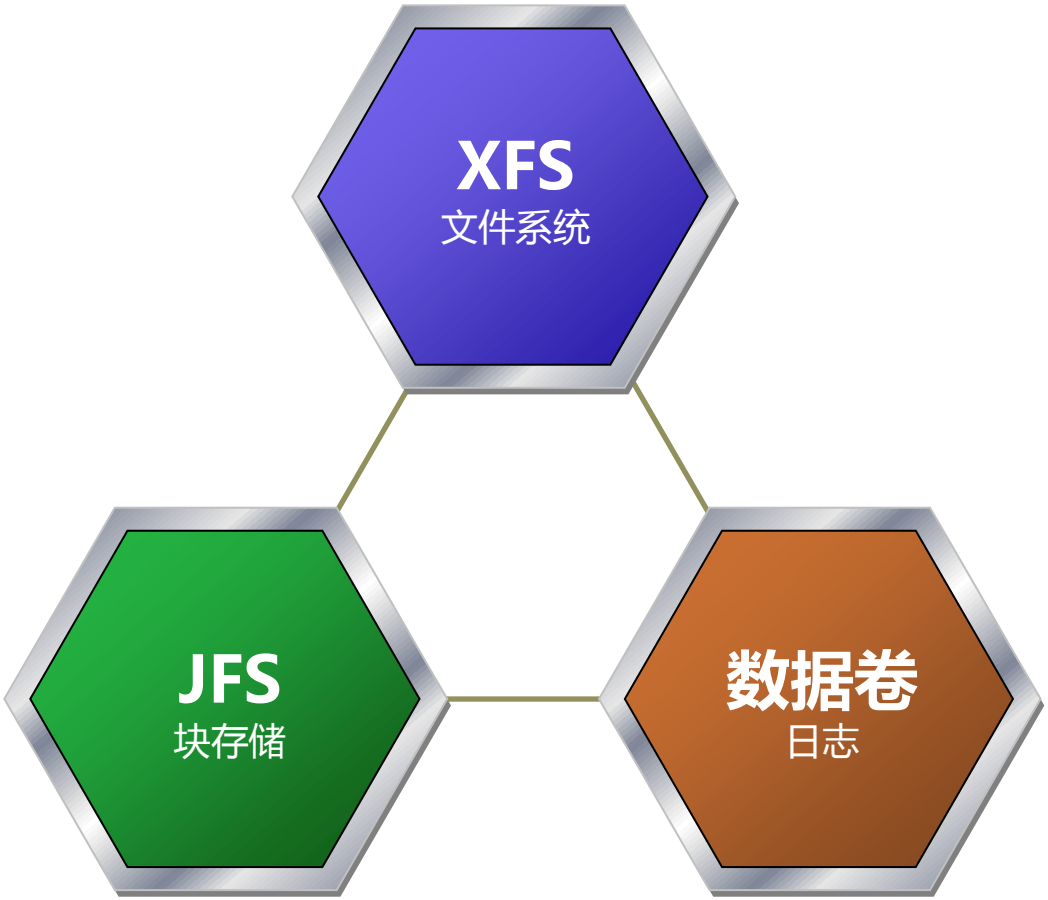

01 成熟度
很成熟，社区非常活跃
OpenStack
02 快速交付
积累了很多经验，快速交付成果
03 一套架构
公有云和私有云一套架构。Windows虚拟机需求

网络（OVS/VLan）

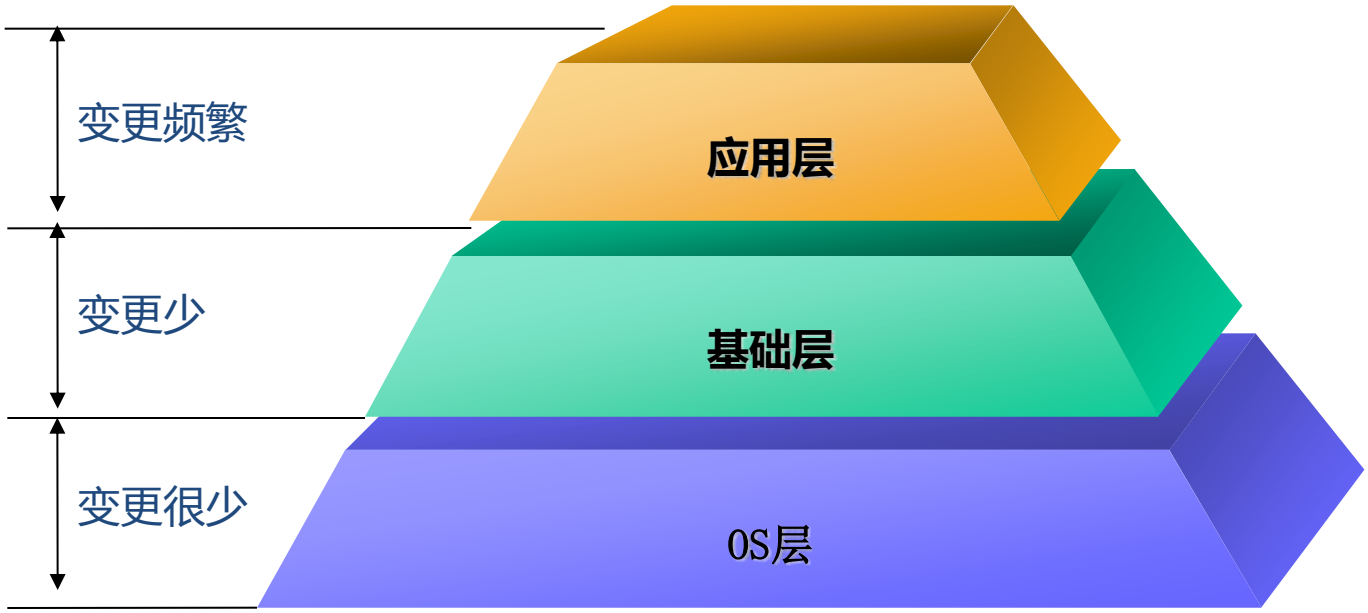


- 为了兼容现在的基础设施系统，满足用户习惯，每个容器都有独立的IP。
- 禁用了Docker网络，采用Neutron集成OVS；
- 优化OVS，提升网络小包延迟，提升性能；

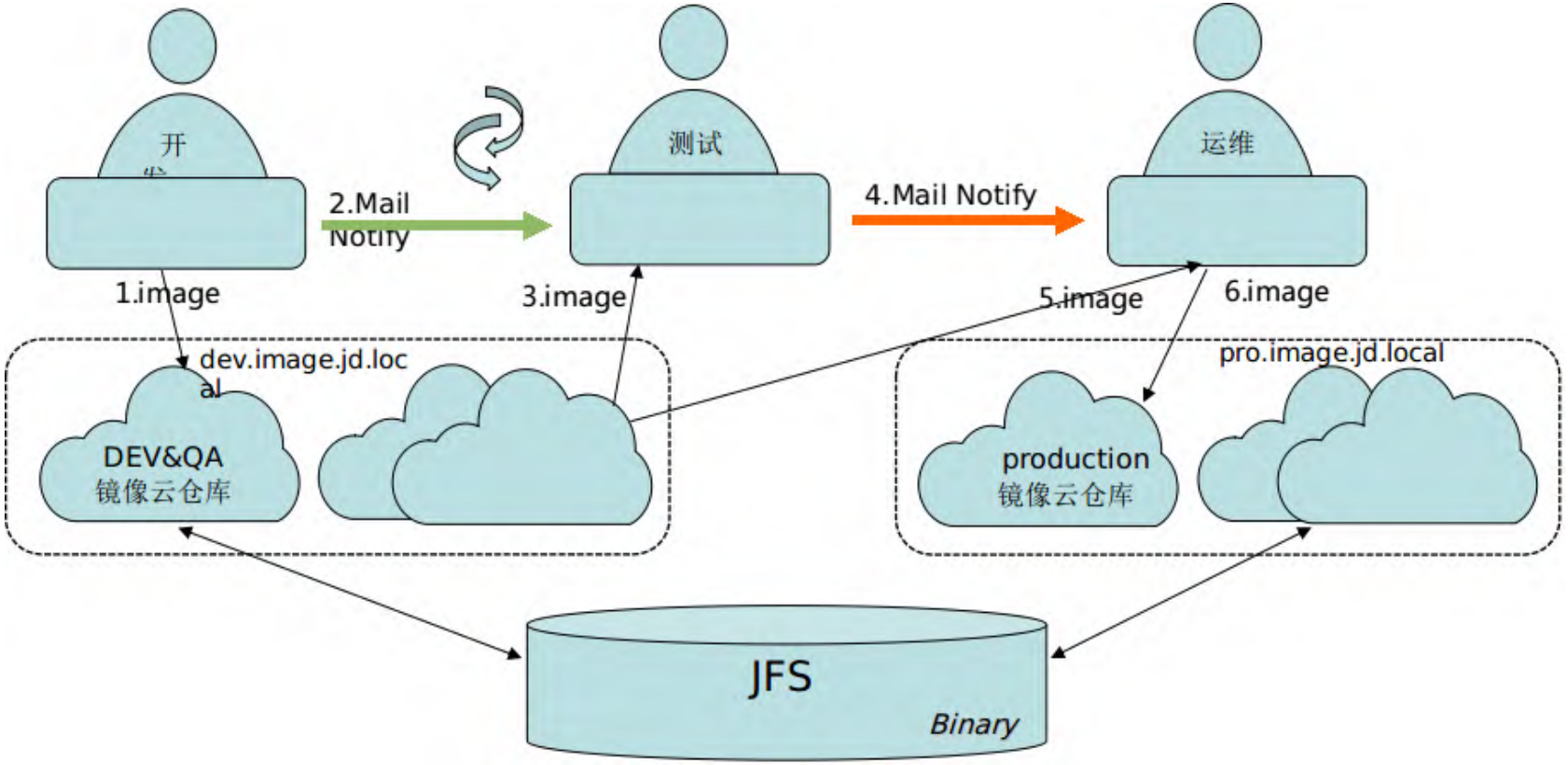
存储



镜像分层合并

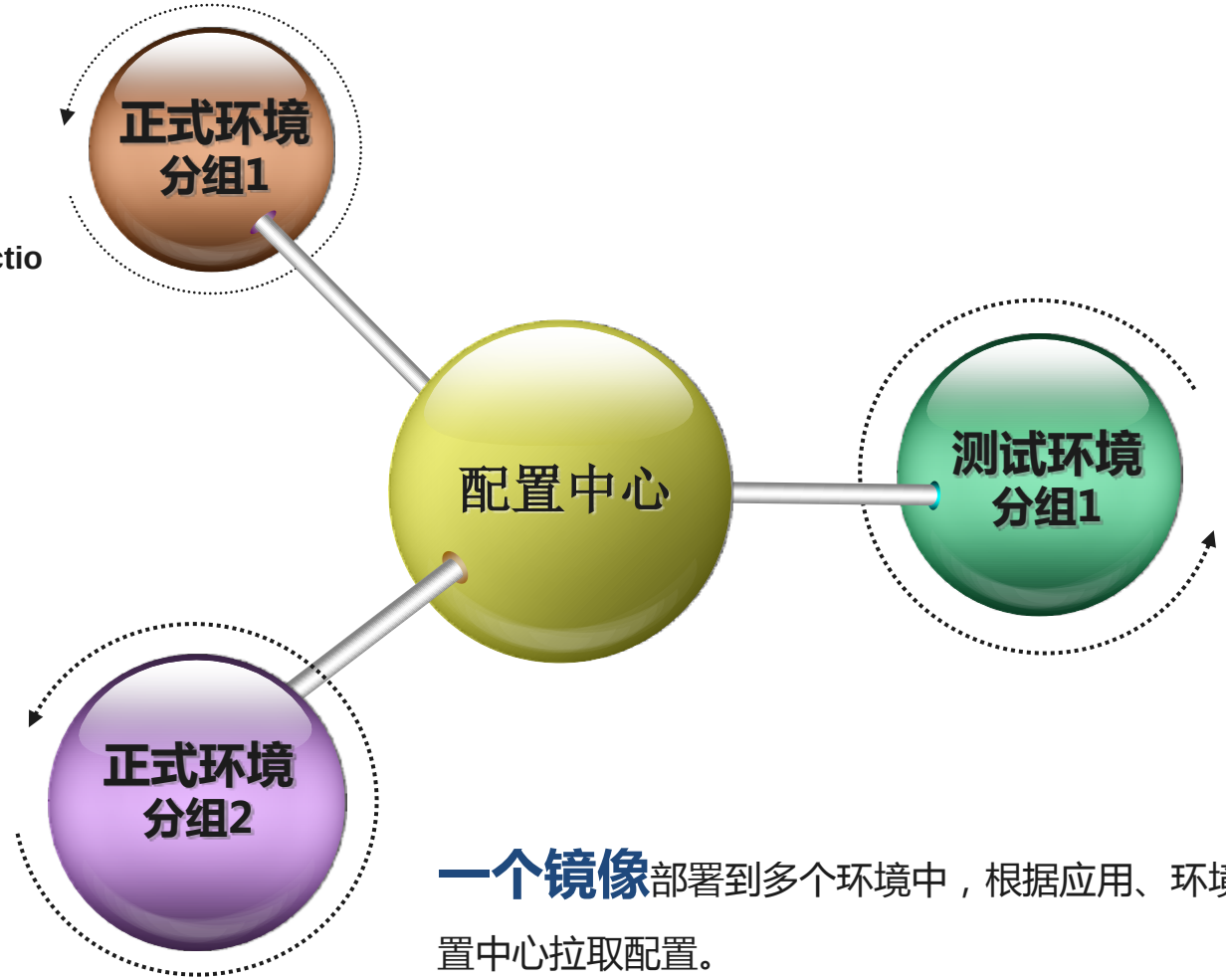


镜像中心


开
测试
运维
2.Mail Notify
4.Mail Notify
1.image
3.image
5.image
6.image
dev.image.jd.local
pro.image.jd.local
DEV&QA
镜像云仓库
production
镜像云仓库
JFS
Binary

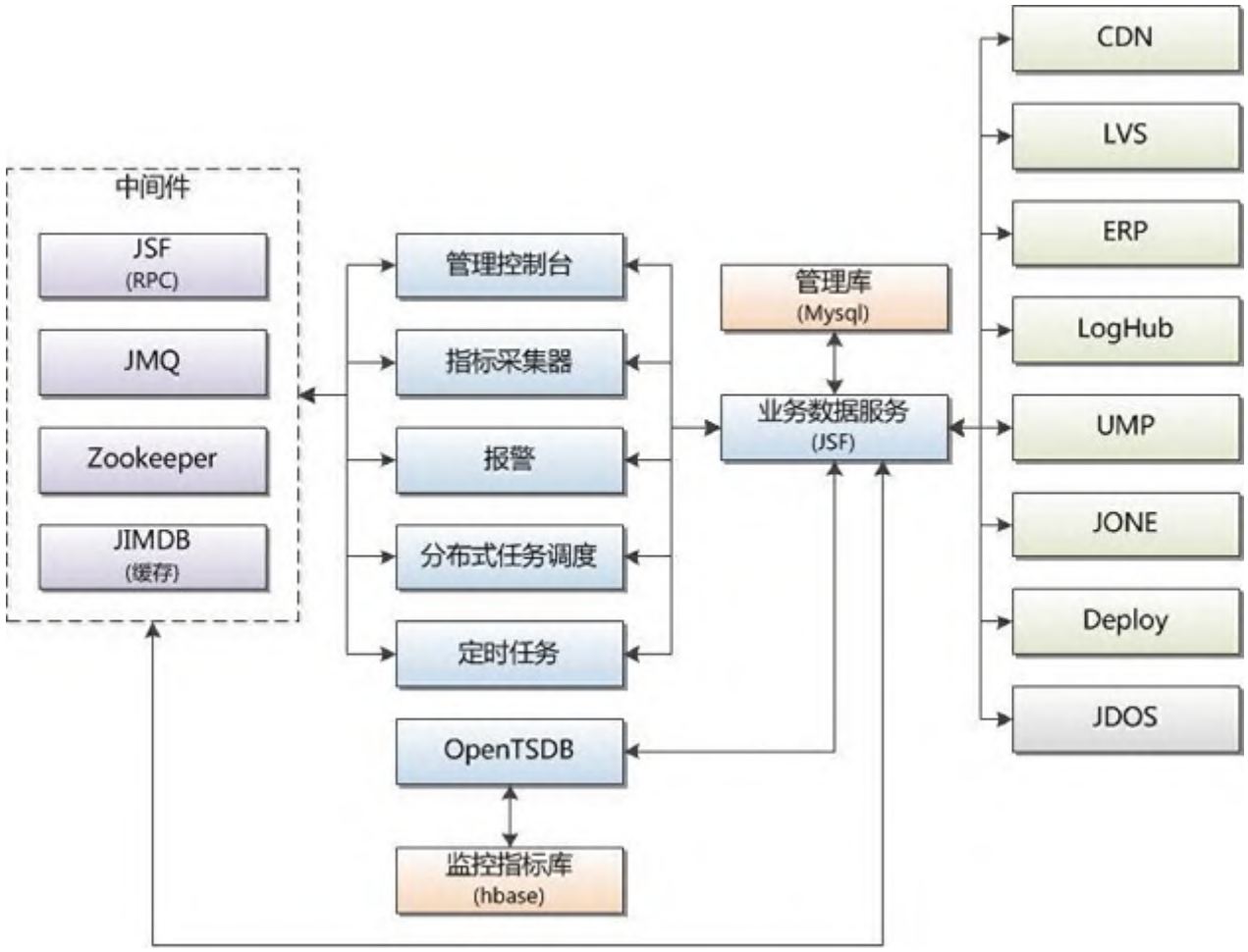
配置中心

- App:CAP
- Group:V2
- Env:Production



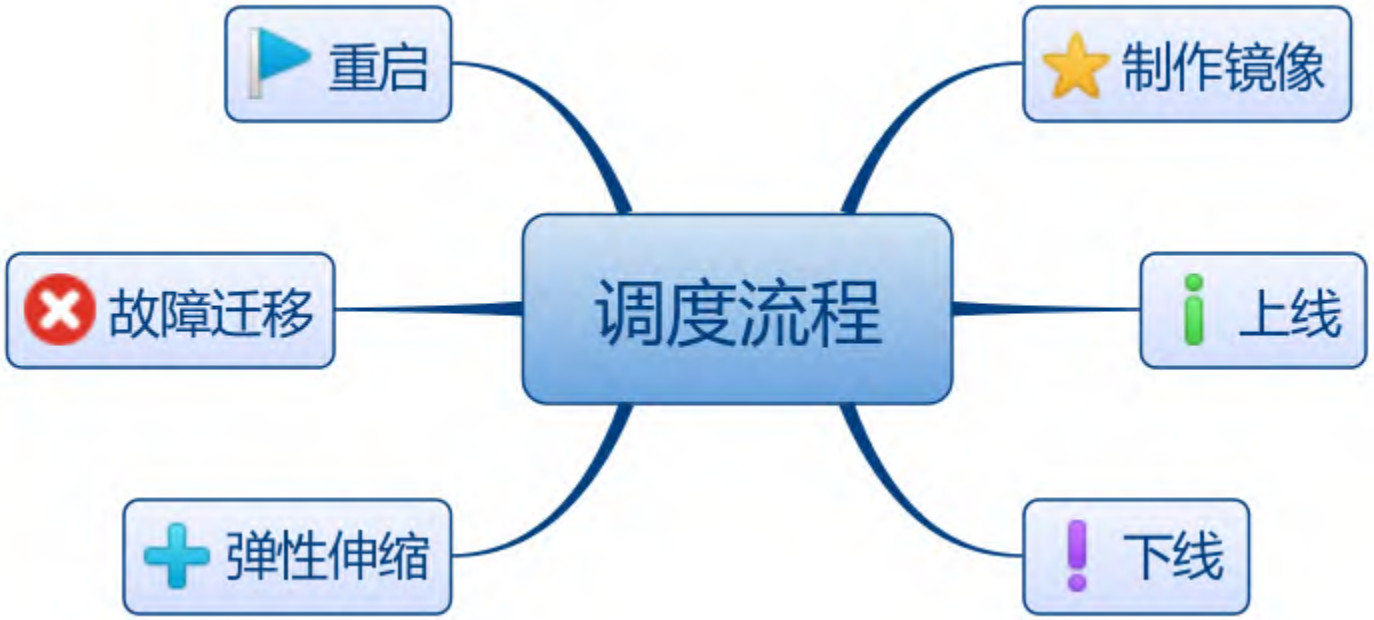
一个镜像部署到多个环境中，根据应用、环境和分组从配置中心拉取配置。

CAP 架构

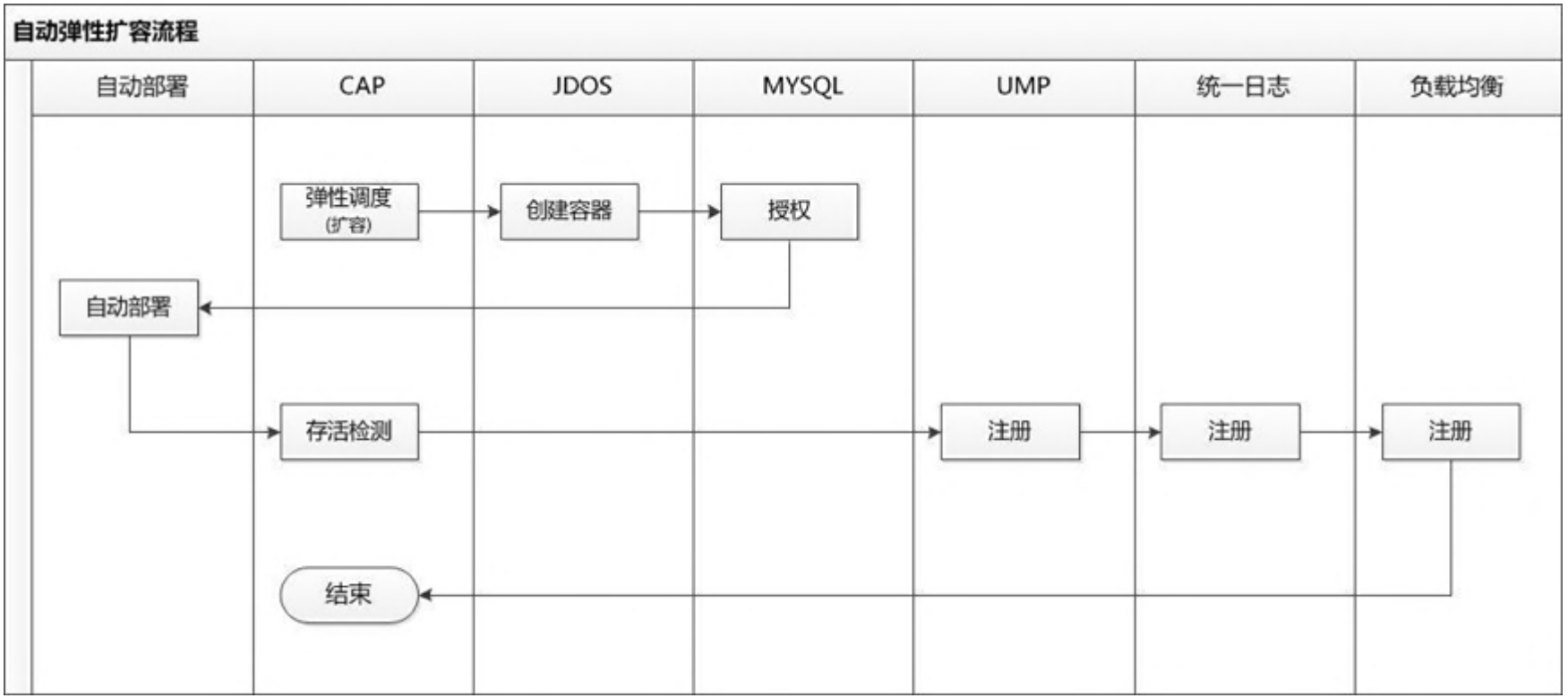


核心是一套工作流，基于Zookeeper分布式调度引擎来实现。能动态注册发现节点；能控制单个节点并发任务数，失败重试次数，确保同一应用互斥任务串行执行。

调度流程

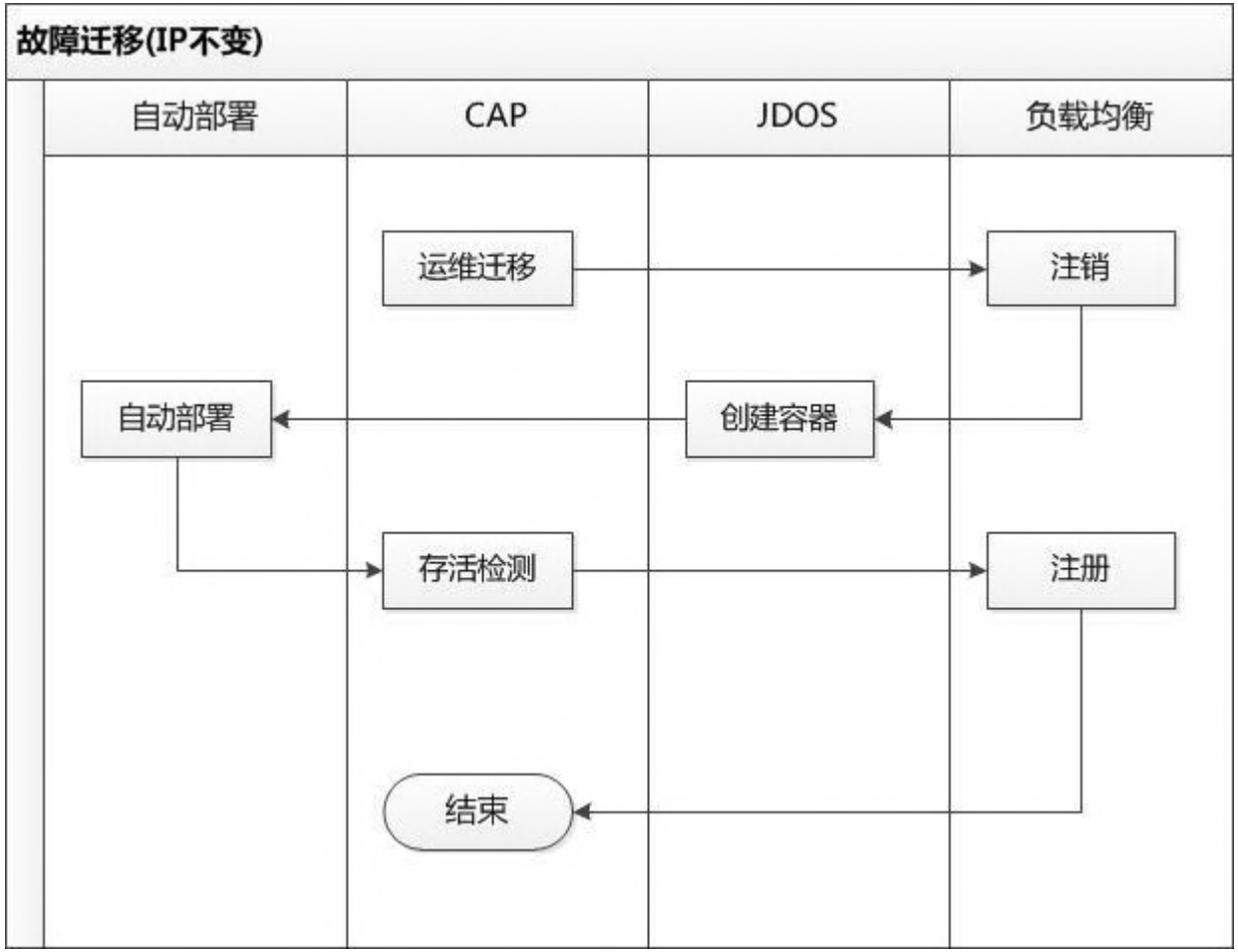


弹性扩容流程



应用在启动之前可能需要数据库授权，启动之后需要挂载VIP，注册统一监控和统一日志。如何能自动发现应用的注册信息，采用了模版方式。应用先申请一个容器，手工注册这些信息，后续的扩容会以该容器为模版来进行自动注册

故障迁移流程



当遇到容器或物理机故障，需要进行快速的迁移，迁移后的容器需要保持原有的IP，避免还要重新申请授权。

弹性调度算法

- 调度单元是应用分组在一个机房的实例。
- 根据应用分组在指定机房的整体负载情况，预测下一时刻负载来进行弹性。

议题

1. 京东容器之路
2. 弹性计算架构
3. **弹性计算应用场景**
4. 自动化运维
5. 数据驱动的精细化运营

应用场景

NGINX

JSF

Worker

redis

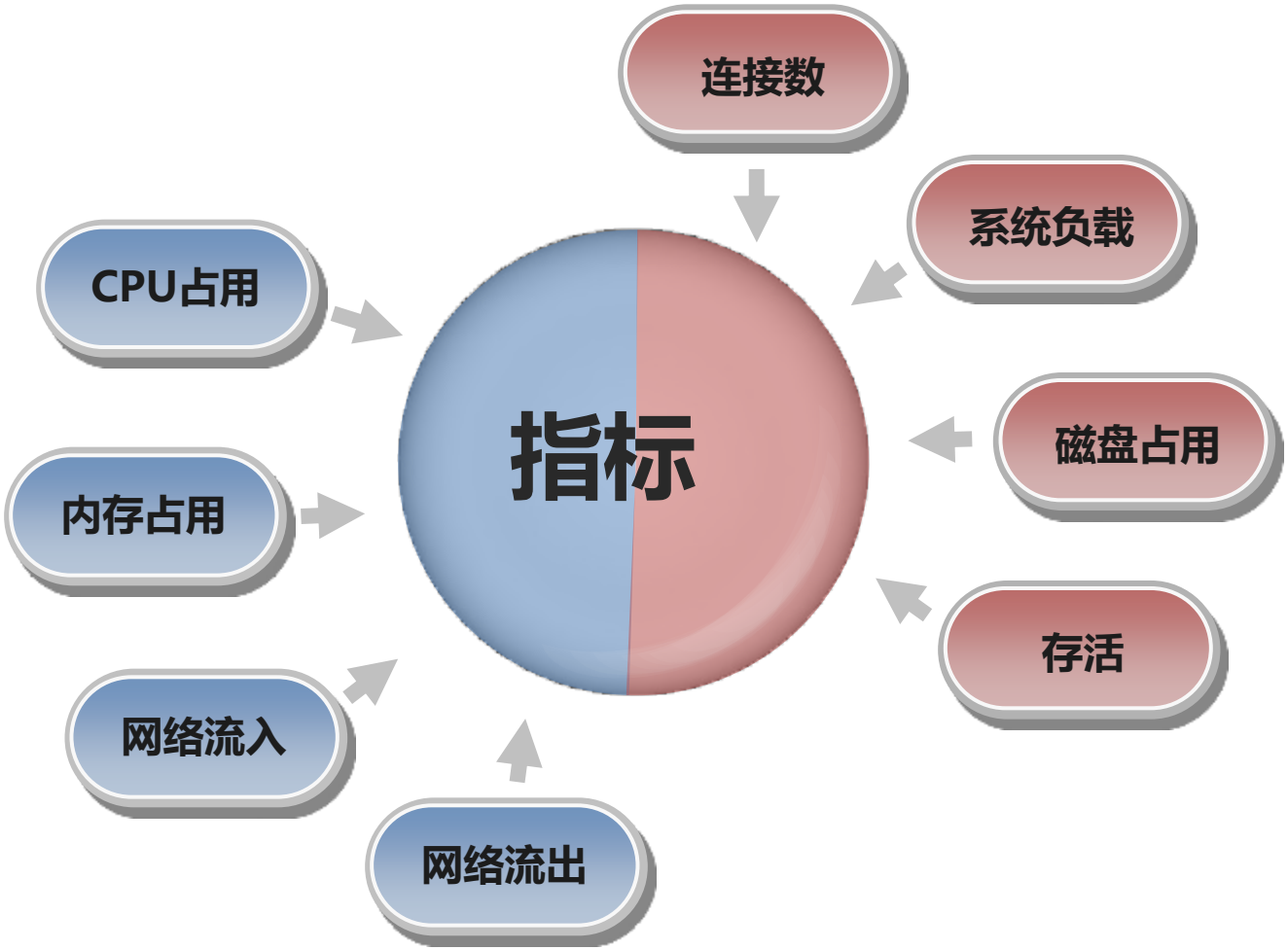
MySQL

- 京东弹性云经过618和双11的大流量考验，新机房以弹性云作为基础架构；
- 核心应用如：网站，交易，订单履约，配送，售后，无线，拍拍，金融，O2O等等平稳运行在容器上

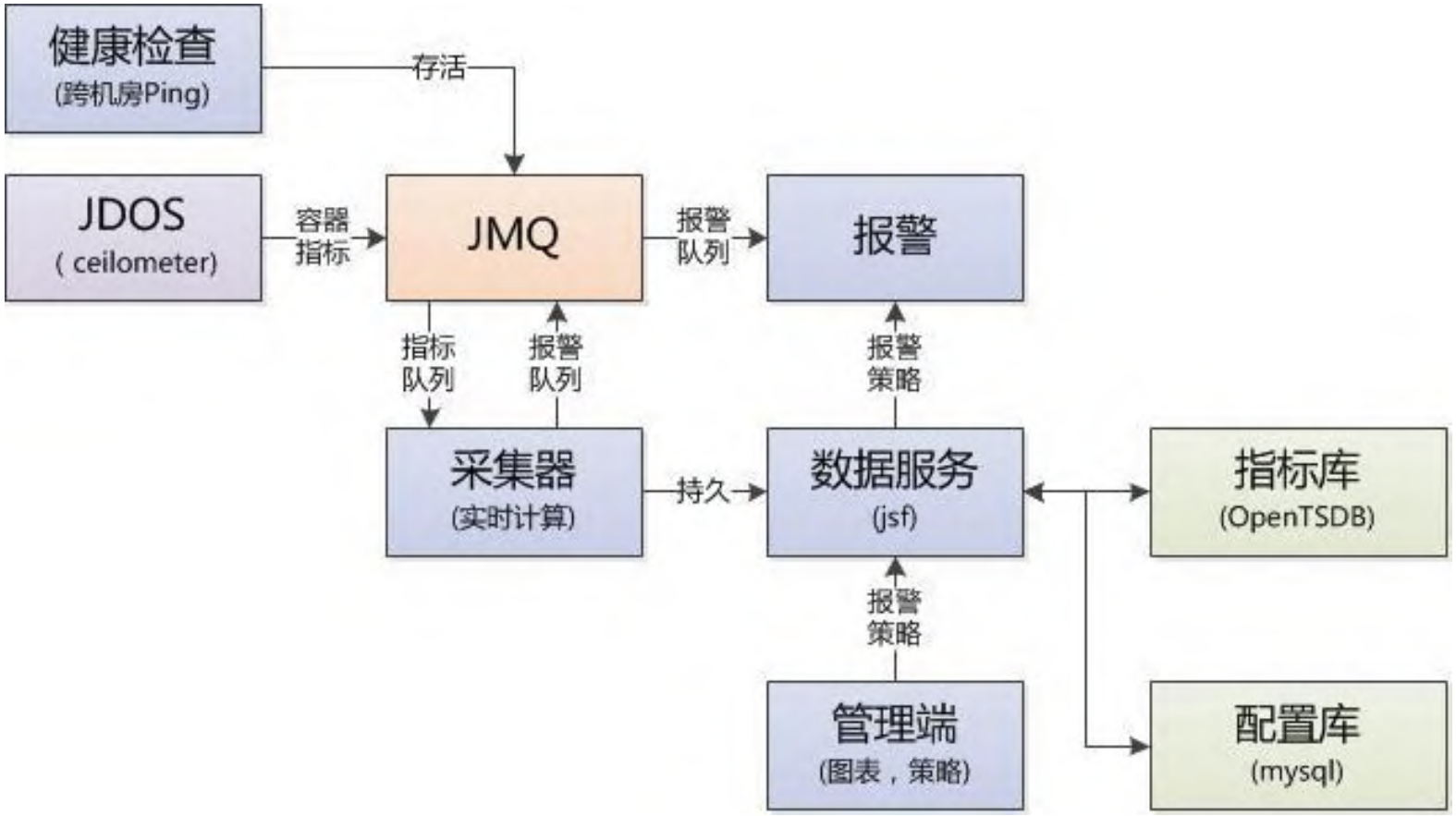
议题

1. 京东容器之路
2. 弹性计算架构
3. 弹性计算应用场景
4. **自动化运维**
5. 数据驱动的精细化运营

系统监控指标



监控架构



■指标数据带有明显的时间特性，每日数据上亿，采用了成熟的OpenTSDB方案。

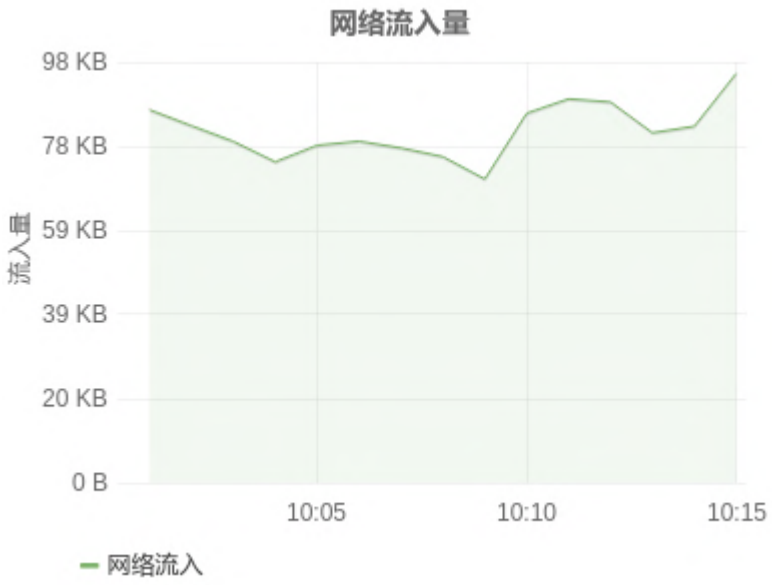
■提供了从应用和实例多个维度查看负载情况，满足用户的需求。

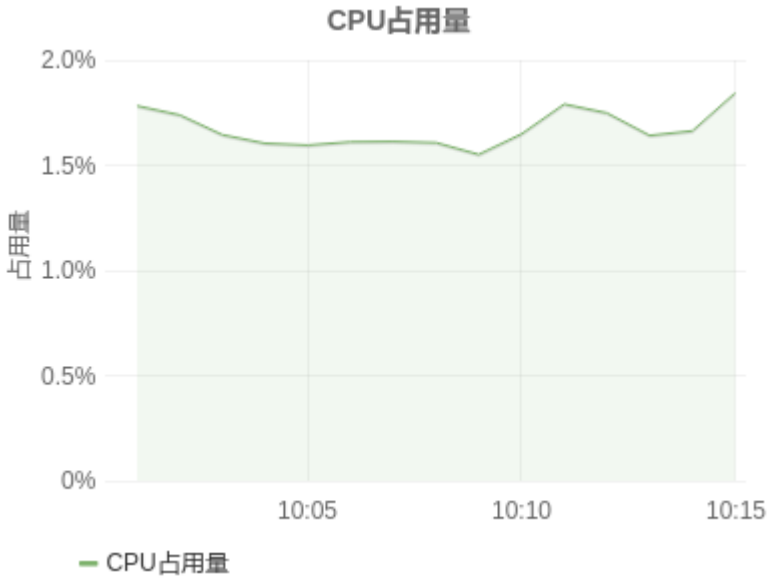
■可以对应用配置警策略，进行短信或邮件报警。

监控页面

警告	严重
2	10

容器IP	主机IP	规格	机房	应用	部门	负责人	CPU	内存	入网	出网	TCP	磁盘	状态
172.19.118.42	172.19.164.39	超配...	永丰	xbopen	京东集团-...	njqiao...							存活
172.19.118.40	172.19.164.39	超配...	永丰	mqsto...	京东集团-...	bjchen...							存活
172.19.118.39	172.19.174.23	标配...	永丰	sk_gaj...	京东集团-...	bjtuh							存活





报警策略

类型	级别	用户	报警方式	应用	阈值	间隔(分钟)	次数	状态	操作
内存占用高	普通	用户	SMS,MAIL	soa.vip	85	0	1	已启用	操作
CPU过载	普通	用户	SMS,MAIL	soa.vip	80	0	1	已启用	操作
CPU过载	普通	用户	SMS,MAIL	vip_php	80	0	1	已启用	操作

系统提供了默认的报警策略。

可以对应用关注的监控指标进行个性化设置。

一键水平扩容

使用水平快照扩容，基于应用的实例，制作快照并进行部署扩容。 需要应用分组提前接入弹性云，并配好相关的VIP路由规则、申请了数据库访问权限。

应用 etms-erp-worker

部署环境 线上正式环境

分组 etms-erp-worker华中组

数量 4

取消 保存

一键垂直扩容

提示：扩容规格cpu，内存，磁盘信息都不能小于现有规格!否则将被过滤。

应用	etms-erp-worker
数据中心	全部
部署环境	全部
分组	选择分组
现有规格	8核/16G内存/50G磁盘
更新规格	8核/32G内存/50G磁盘

取消　保存

一键水平缩容

注意！目前的策略是随机执行容器进行缩容

应用 etms-erp-worker

部署环境 线上正式环境

分组 etms-erp-worker全国组

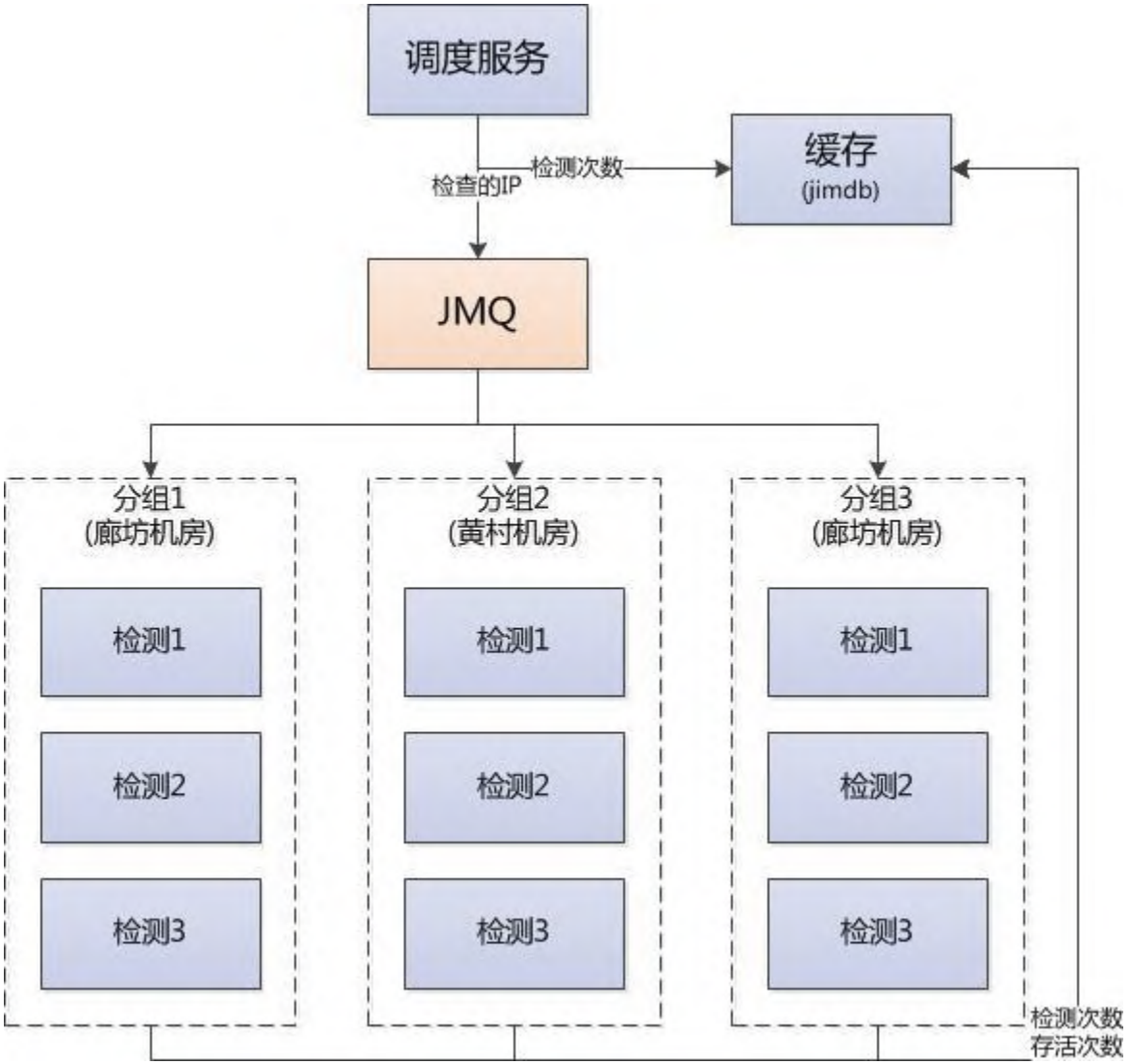
数量 1

取消 保存

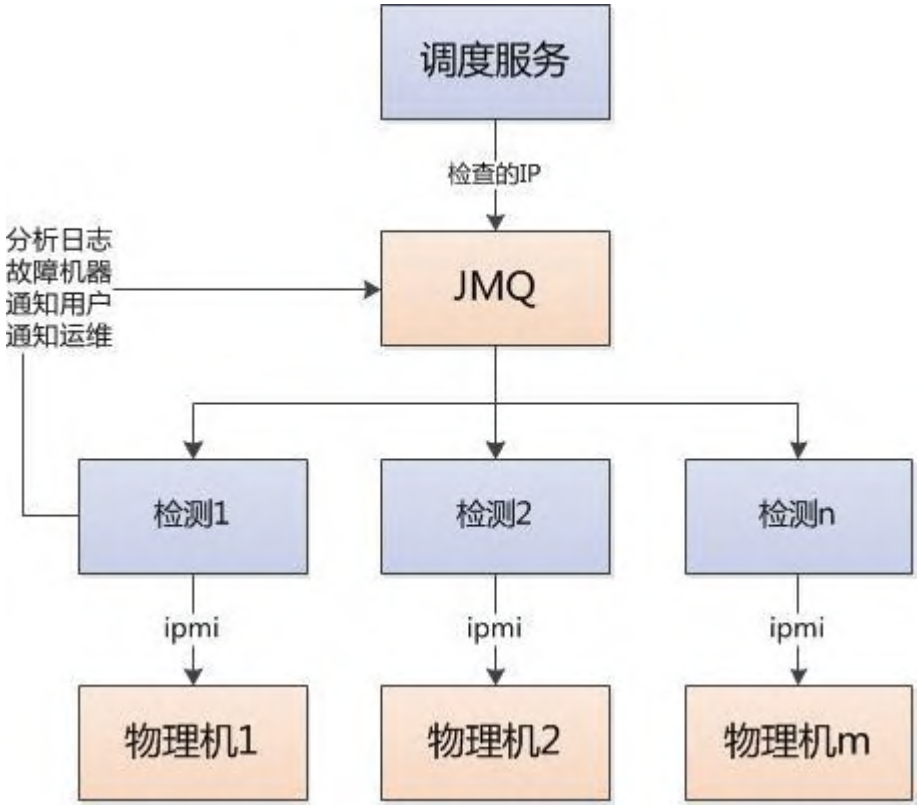
一键垂直缩容


应用
etms-erp-worker
数据中心
廊坊1
部署环境
线上正式环境
分组
etms-erp-worker全国组
现有规格
8核/16G内存/50G磁盘
更新规格
4核/16G内存/50G磁盘
是否强制
否
是
警告！强制执行缩容将导致内存和磁盘也会减少，可能导致应用不可用!
取消
保存

宕机探测架构


调度服务
检查的IP
检测次数
缓存
(jimdb)
JMQ
分组1
(廊坊机房)
分组2
(黄村机房)
分组3
(廊坊机房)
检测1
检测2
检测3
检测1
检测2
检测3
检测1
检测2
检测3
检测次数
存活次数

硬件故障探测


调度服务
检查的IP
分析日志
故障机器
通知用户
通知运维
JMQ
检测1
检测2
检测n
ipmi
ipmi
ipmi
物理机1
物理机2
物理机m

故障通知

IP	原因	申请人	恢复人	提交时间	修改时间	状态	操作
172.20.100.60	变更物理机磁盘故...	姬.....	姬.....	2016-03-04 13:52:24	2016-03-04 13:53:00	运维邮件已发	操作
172.20.100.27	物理机内存故障，...	姬.....	姬.....	2016-03-04 10:55:09	2016-03-04 10:56:00	运维邮件已发	操作
172.20.102.52	物理机内存故障，...	姬.....	姬.....	2016-03-03 10:06:39	2016-03-03 10:07:00	运维邮件已发	操作
172.22.85.74	电源故障，服务器...	姬.....	姬.....	2016-03-02 17:36:51	2016-03-02 17:37:00	运维邮件已发	操作
172.20.78.102	物理机内存故障，...	姬.....	姬.....	2016-03-02 11:19:18	2016-03-04 13:46:00	恢复邮件已发	操作
172.20.119.12	物理机内存故障，...	姬.....	姬.....	2016-03-02 11:19:18	2016-03-04 13:46:00	恢复邮件已发	操作
172.22.83.99	物理机 PCIE总线故...	姬.....	姬.....	2016-03-02 10:49:10	2016-03-03 08:56:00	恢复邮件已发	操作
172.20.118.22	物理机内存故障，...	姬.....	姬.....	2016-03-01 17:31:05	2016-03-04 13:46:00	恢复邮件已发	操作
172.28.129.56	CPU故障导致物理...	姬.....	姬.....	2016-03-01 10:21:40	2016-03-02 17:57:00	恢复邮件已发	操作
172.20.90.21	物理机内存故障，...	姬.....	姬.....	2016-03-01 09:40:16	2016-03-04 13:46:00	恢复邮件已发	操作

应用部署巡检

- 定期巡检应用容器部署情况，邮件报告；

超载 01

- 单个机房部署过多
- 单个交换机部署过多
- 单个物理机部署过多

未部署 02

- 申请的容器没有使用

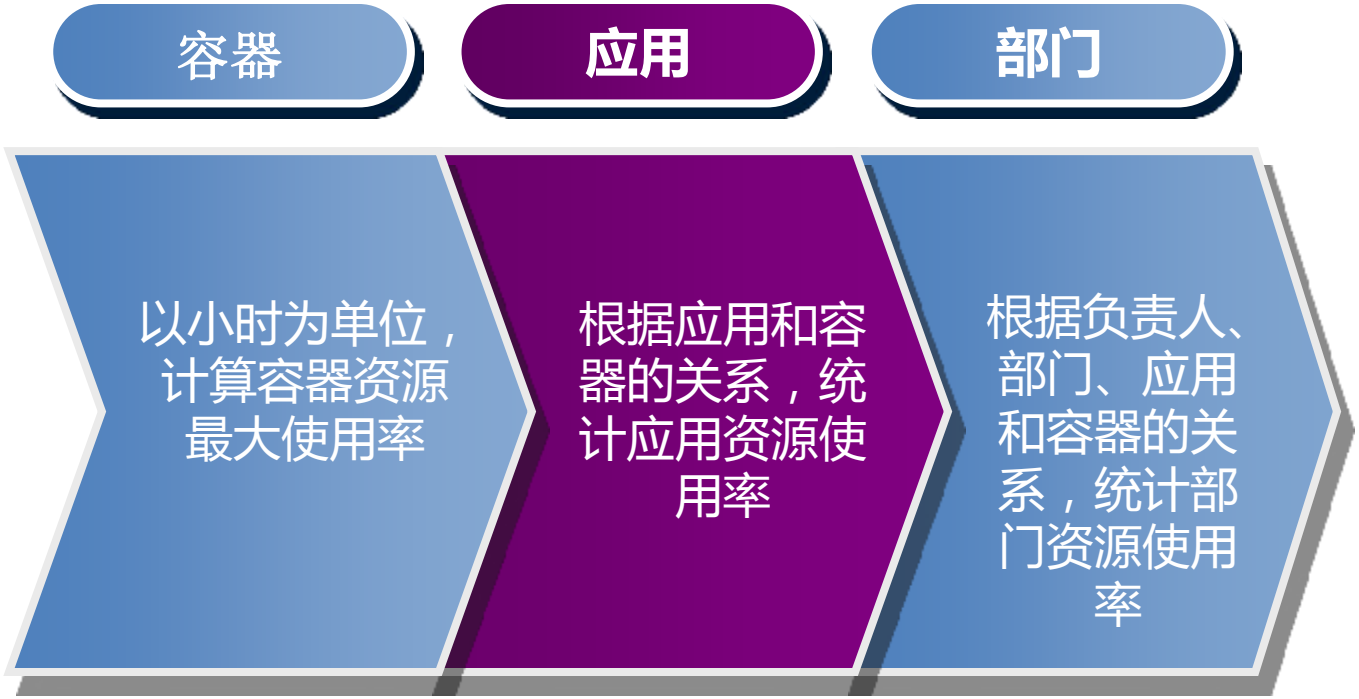
规格不一致 03

- 容器规格不均匀，可能造成流量负载不均匀

议题

1. 京东容器之路
2. 弹性计算架构
3. 弹性计算应用场景
4. 自动化运维
5. **数据驱动的精细化运营**

资源利用率


容器
应用
部门
以小时为单位，计算容器资源最大使用率
根据应用和容器的关系，统计应用资源使用率
根据负责人、部门、应用和容器的关系，统计部门资源使用率

容器资源利用率

应用名称	代码	部门	域名	负责人	IP	主机	分组	数据中心	ZONE	规格	使用率	CPU	内存	入网	出网	TCP	磁盘
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	30.66%	8.06%	30.66%	56844...	508128	25	13%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	17.8%	2.45%	17.8%	7849.16	14896....	28	15%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	17.32%	2.24%	17.32%	8197.73	19968....	28	32%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	17.94%	2.18%	17.94%	11292....	12842....	28	15%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	17.22%	2.1%	17.22%	7098.3	11652....	29	51%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	40.34%	5.77%	40.34%	96237...	15652...	113	54%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	48.47%	25.76%	48.47%	44550...	36142...	81	36%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	73.63%	73.63%	50.34%	13050...	83189...	128	81%
jone_test	jon...	京...	jon...	bjlifa...	17...	17...	ces	永丰	nova	[4核/8G...	4.44%	1.27%	4.44%	6343.83	10390....	4	32%
弹性计...	ca...	京...	ca...	bjhe...	17...	17...	正式...	永丰	nova	[4核/8G...	17.07%	2.54%	17.07%	21600...	18464...	96	13%

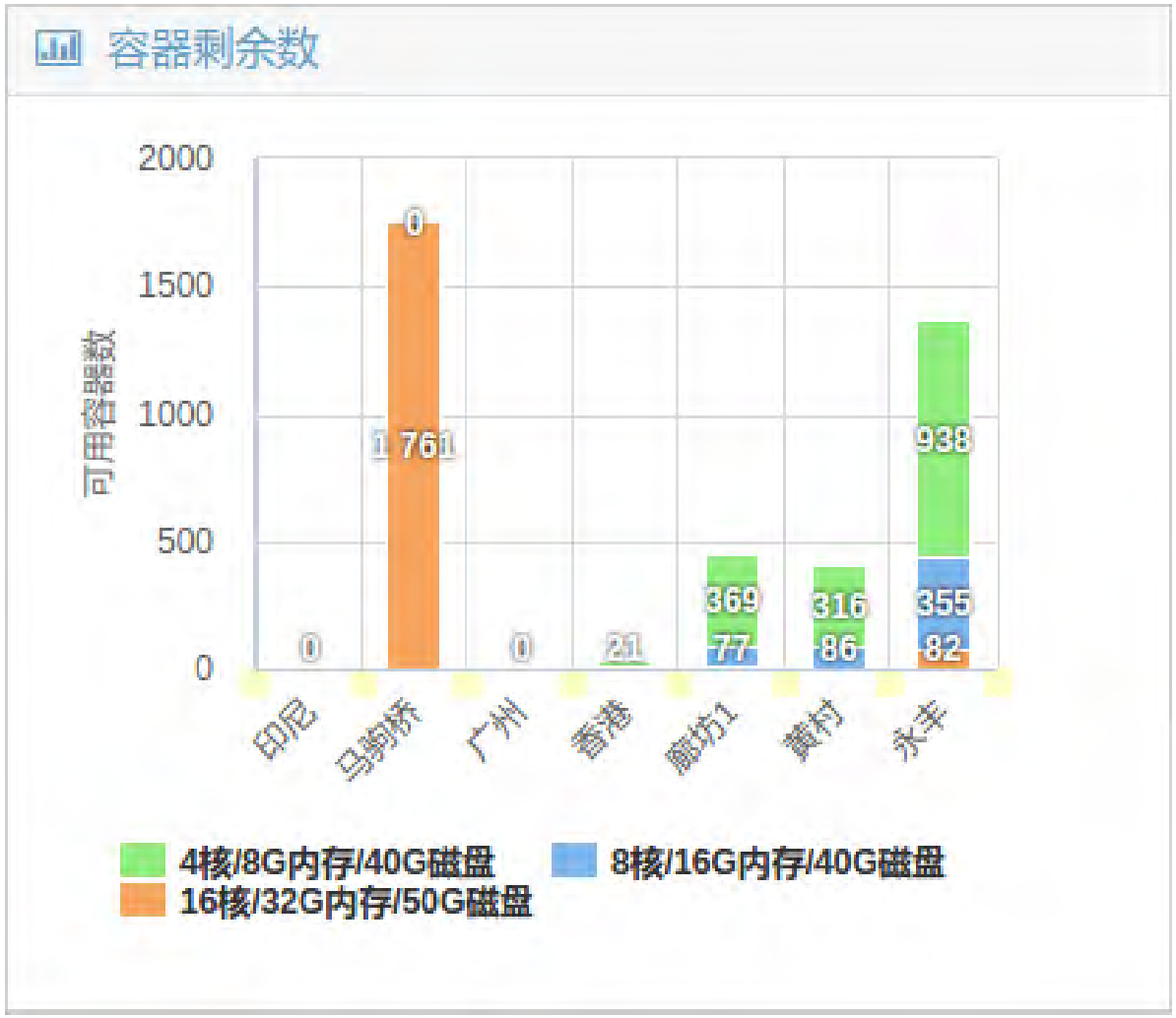
应用资源利用率

应用代码	应用名称	部门	负责	容器数量	核数	CPU	内存	入网	出网	TCP	磁盘
tmspda...	青龙配...	京东集团-京东商城...	苗...	3	12	1.07%	9.68%	19332.98	11512.46	337	8%
dmsrec...	B商家...	京东集团-京东商城...	王...	4	16	3.14%	25.37%	146312.6	207765.91	354	14%
contrac...	B商家...	京东集团-京东商城...	王...	6	24	0.96%	25.71%	1317.01	1165.31	134	11%
rec.etms	B商家...	京东集团-京东商城...	王...	2	8	7.29%	35.36%	643705.86	423757.01	224	12%
dms-re...	B商家...	京东集团-京东商城...	王...	4	16	1.95%	23.97%	122548.67	121371.44	309	8%
sep.etm...	sep.et...	京东集团-京东商城...	吴...	4	16	75.86%	26.93%	2205.73	1719.52	273	10%
presep...	presep...	京东集团-京东商城...	吴...	6	24	63.49%	21.59%	958730.75	3700513.95	422	73%
man.m...	景点运...	京东集团-京东商城...	崔...	2	16	0.4%	10.16%	35830.58	18809.96	150	11%
chat.er...	老版在...	京东集团-CTO体系...	黄...	5	10	38.47%	33.39%	52094.4	109313.6	256	4%
jshop	JSHOP...	京东集团-CTO体系...	赵...	37	148	2.02%	24.03%	220877.41	119840.58	664	10%

部门资源利用率

部门	容器数量	资源使用率	CPU	内存	入网	出网	TCP	磁盘
京东集团...	427	21.93%	8.91%	21.93%	3258337.13	3059869.06	2276	25%
京东集团...	75	18.87%	9%	18.87%	113675463....	63143689.08	436	77%
京东集团...	519	22.46%	15.51%	22.46%	1276289.66	981562.81	2288	88%
京东集团...	428	21.02%	12.39%	21.02%	63329789.73	6239280.21	1548	89%
京东集团...	14	17.29%	10.17%	17.29%	2553944.4	384858.66	107	65%
京东集团...	277	22.62%	2.57%	22.62%	5598788.73	3458102.3	2038	74%
京东集团...	160	20.94%	11.09%	20.94%	3196754.2	1221896.83	601	33%
京东集团...	380	31.78%	16.66%	31.78%	80524159.4	25227123.41	1766	91%
京东集团...	1793	19.31%	7.23%	19.31%	21229138.43	31931908.78	9760	100%
京东集团...	228	11.92%	4.68%	11.92%	822996.3	500376.78	1862	82%

资源剩余情况


容器剩余数
2000
1500
1000
500
0
可用容器数
0
1761
0
21
369
77
316
86
355
82
938
印尼
马驹桥
广州
香港
廊坊1
廊坊
永丰
4核/8G内存/40G磁盘
8核/16G内存/40G磁盘
16核/32G内存/50G磁盘

配额管理

部门	父部门	部门级别	数据中心	总配额	已用配额	可用配额	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	黄村	84	84	0	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	廊坊1	720	720	0	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	香港	0	0	0	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	广州	0	0	0	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	马驹桥	64	40	24	操作
京东集团-CTO体系-CTO办公室	京东集团-CTO体系	一级部门	印尼	0	0	0	操作

实践经验

- 无状态，同时对磁盘IO要求不高的应用，很适合部署到弹性云；
- 微服务应用由于能自动服务注册发现，辅助均衡，非常适合部署到弹性云
- 推荐万兆网络和网卡，避免网络共享出现资源竞争；
- 稳定的操作系统版本；
- 推荐高配置物理机，合理得CPU和内存比，便于充分利用资源；
- 采购高质量的交换机和物理机；

谢谢