京东弹性计算实践

京东/云平台/系统技术部 何小锋

JD.京东.COM

自我介绍

JD.COM 京东

姓名：何小锋

部门：京东/云平台/系统技术部

职位：高级架构师

邮箱：hexiaofeng@jd.com

电话：13910526009

个人介绍：

拥有17年的研发经验，喜欢技术，追求卓越。

2011年加入京东，担任了京东两届架构委员会常委。

目前在京东云平台系统技术部，负责弹性计算和分布式消息平台建设。

议题

JD.COM 京东

1. 京东弹性计算之路
2. 京东弹性计算架构
3. 京东弹性调度策略及算法
4. 京东弹性调度流程
5. 京东弹性计算应用场景

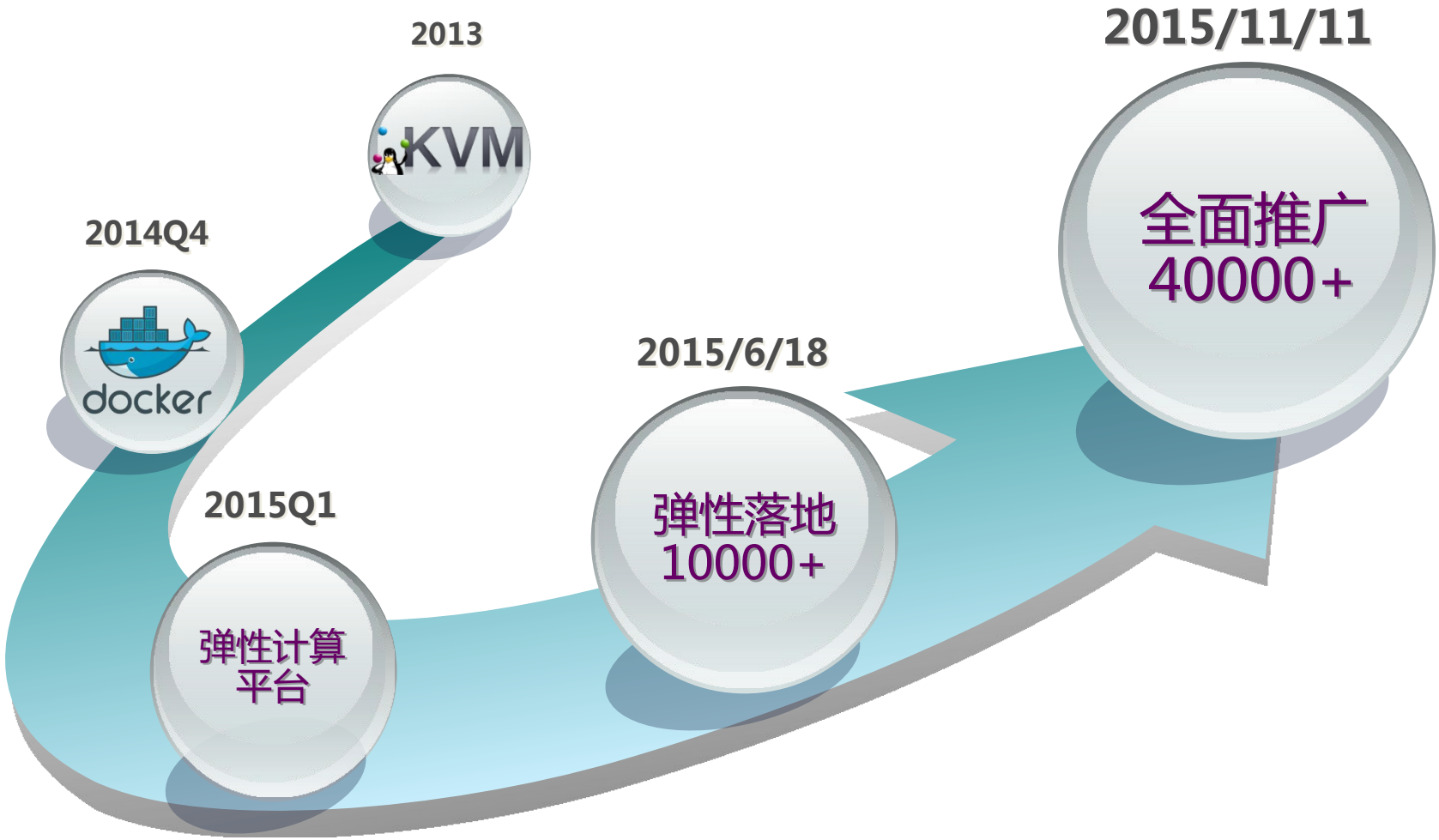
面临得挑战

JD.COM 京东

- 随着京东业务的发展，应用越来越多，更新迭代很快，目前交付效率不高。
- 每年物理机的增速快，运维成本高；
- 为了满足618、双11和秒杀活动的性能需求，需要提前准备大量服务器，资源浪费严重。
- 当监控到应用负载高，需要扩容的时候，部署新实例流程复杂，扩容慢。
- 应用部署拓扑复杂，多环境部署，多机房部署，为了减少网络调用，多个依赖应用混合部署在一起；

京东弹性计算之路

JD.COM 京东



从KVM迁移到Docker

JD.COM 京东

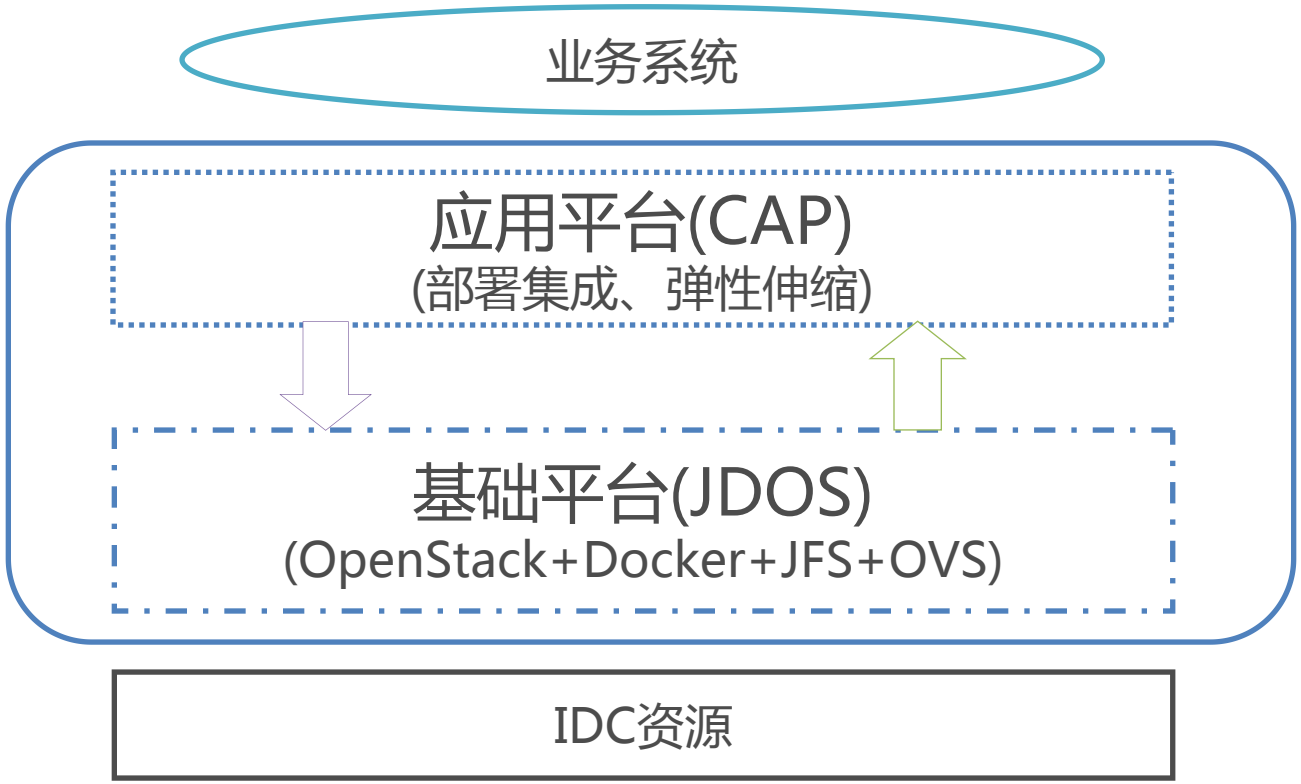
议题

JD.COM 京东

1. 京东弹性计算之路
2. **京东弹性计算架构**
3. 京东弹性调度策略及算法
4. 京东弹性调度流程
5. 京东弹性计算应用场景

京东弹性计算架构

JD.COM 京东



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；

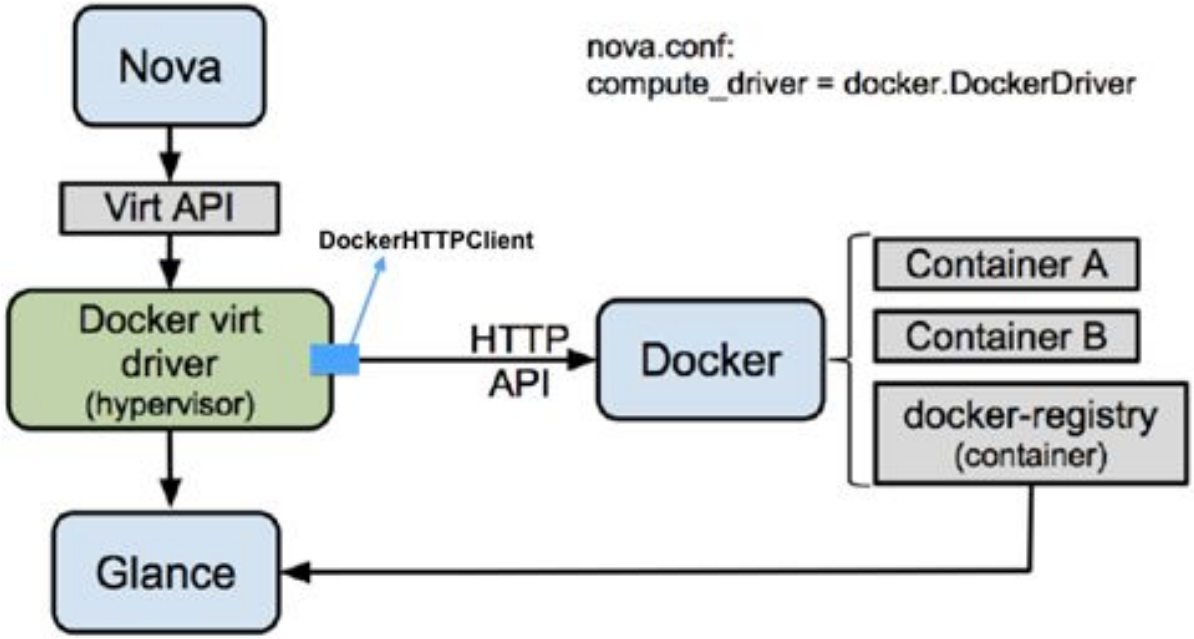
■CAP负责应用治理、部署、监控报警、资源利用率统计、手动扩容和缩容、自动弹性伸缩。

OpenStack+Docker

JD.COM 京东

业界还没有成熟的Docker集群管理系统，OpenStack经过几年的发展，已经相对成熟，JDOS沿用了OpenStack来管理Docker。

采用Nova Docker Driver方案来集成，把Docker作为一种新的Hypervisor来处理，这样可以使用OpenStack中的所有服务，包括使用Nova Scheduler来做资源调度，集成Neutron来管理Docker网络，支持多租户和资源隔离等等



JDOS网络及存储

JD.COM 京东

- 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。
- 当前版本Docker自带的网络功能无法满足现有需求，禁用了Docker网络，采用Neutron集成OVS。
- 京东推广了微服务，大量的调用传递的数据包较小，优化OVS转发层，显著提升网络小包延迟，提升比率大约有20%；
- 目前版本的DeviceMapper还是实验性产品，不稳定，生产环境使用的是XFS文件系统。
- 用户经常有查询日志的需求，通过挂载数据盘，把日志保留到宿主机，避免容器销毁后，日志丢失。
- 同时集成了京东自主研发的JFS作为块存储。

JDOS镜像

JD.COM 京东

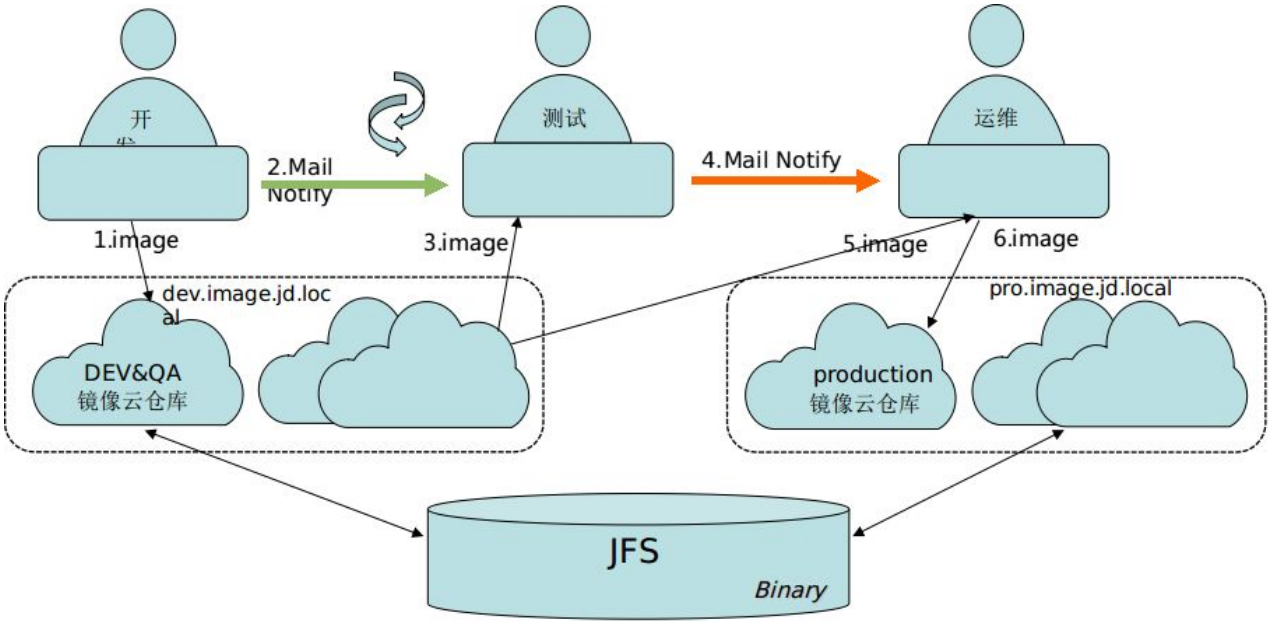
京东应用发布更新很快，部署复杂。通过在一台容器上做好镜像，可以快速的分发到线上，能极大的提高效率。

■为了减少镜像的大小，镜像分层为OS层、基础层和应用层，支持镜像合并。

■为了加快分发的速度，采用镜像预分发技术，OS层和基础层提前分发，应用层会在创建容器的时候实时拉取，支持镜像数据文件分段并行下载。

■根据公司项目管控的需要，分别构建开发测试环境和线上正式环境镜像中心

■不同的环境有不同的配置，提供了统一的配置中心来实现配置的动态修改。

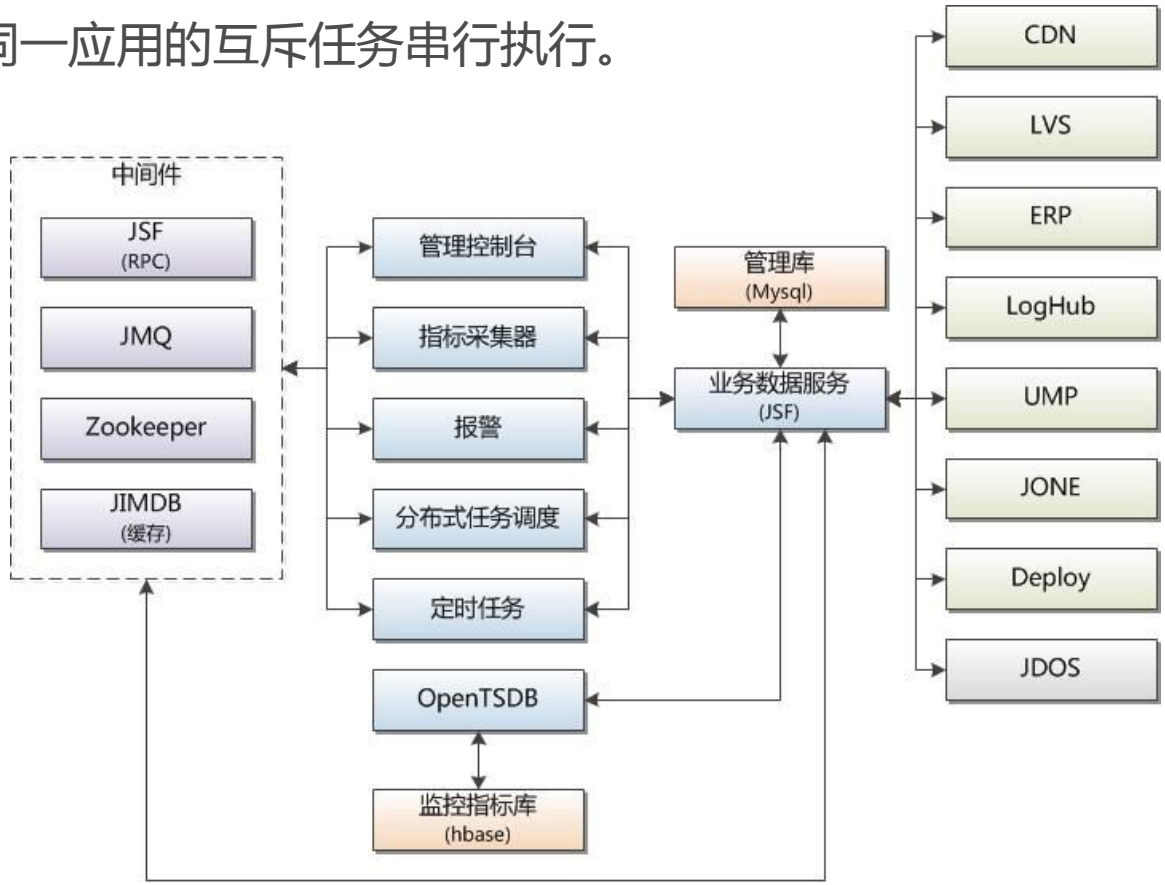


CAP—架构

JD.COM 京东

CAP集成现有基础系统，实现镜像制作，部署流程，手动扩容和缩容、自动弹性伸缩，应用治理、监控报警、资源利用率统计等等功能。

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，便于动态扩容；能控制单个节点的并发任务数，任务失败后的重试次数，同时确保同一应用的互斥任务串行执行。



CAP—监控方案

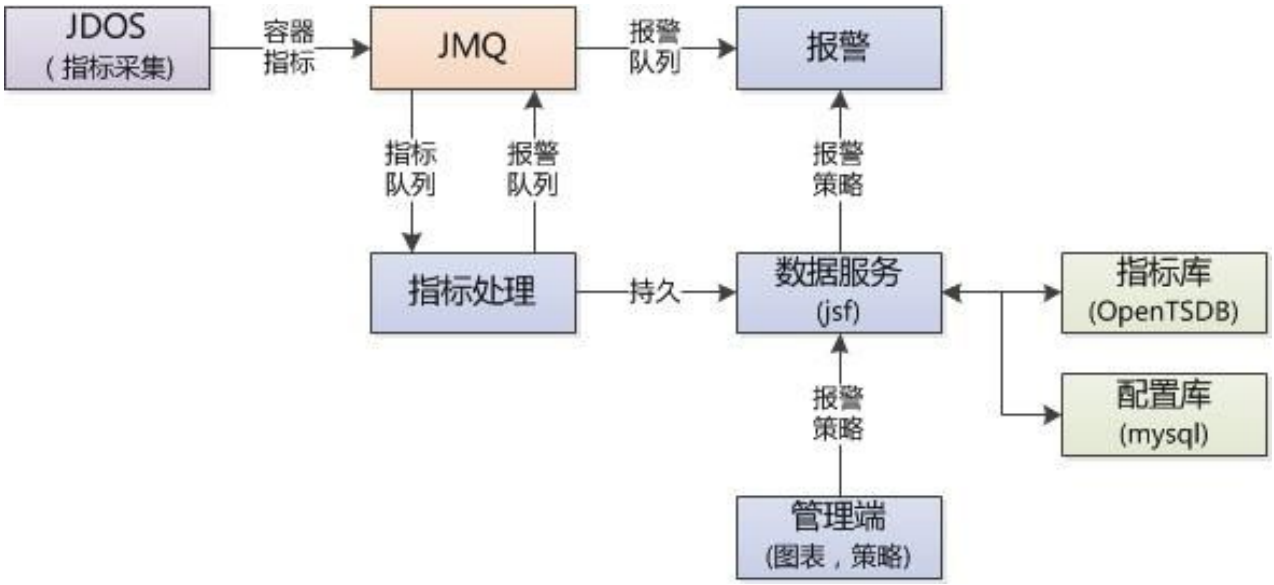
JD.COM 京东

监控报警、弹性伸缩、资源利用率等等都依赖于容器指标，要求数据实时准确。用户期望容器的监控指标和物理机的一致，包括CPU，内存，网络流量，系统负载，连接数等等，自研采集程序补充指标。

■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。

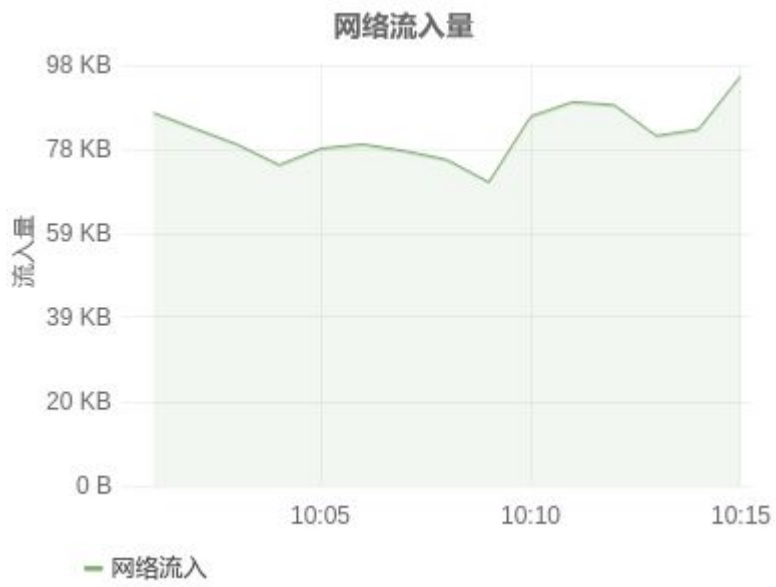


CAP—监控页面

JD.COM 京东

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



议题

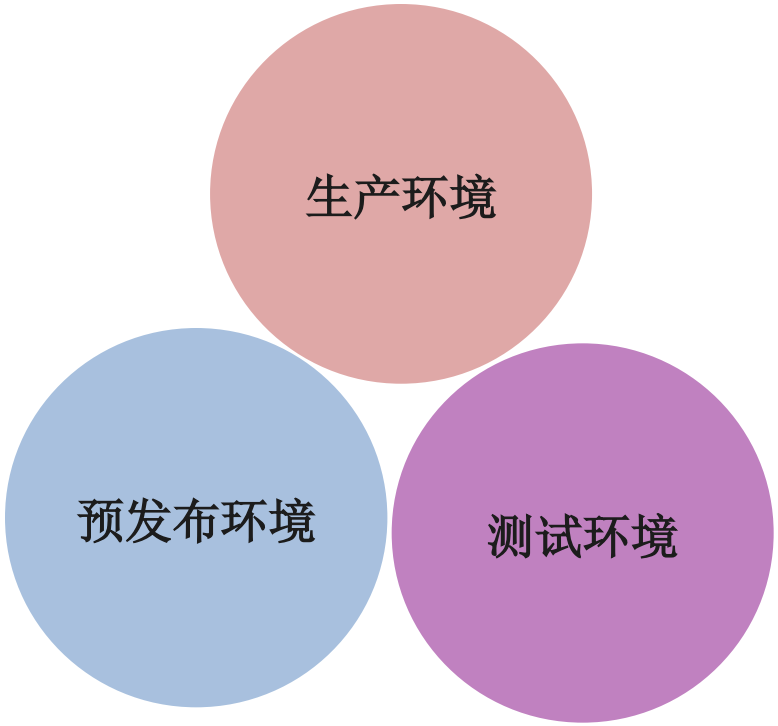
JD.COM 京东

1. 京东弹性计算之路
2. 京东弹性计算架构
3. **京东弹性调度策略及算法**
4. 京东弹性调度流程
5. 京东弹性计算应用场景

多个部署环境

JD.COM 京东

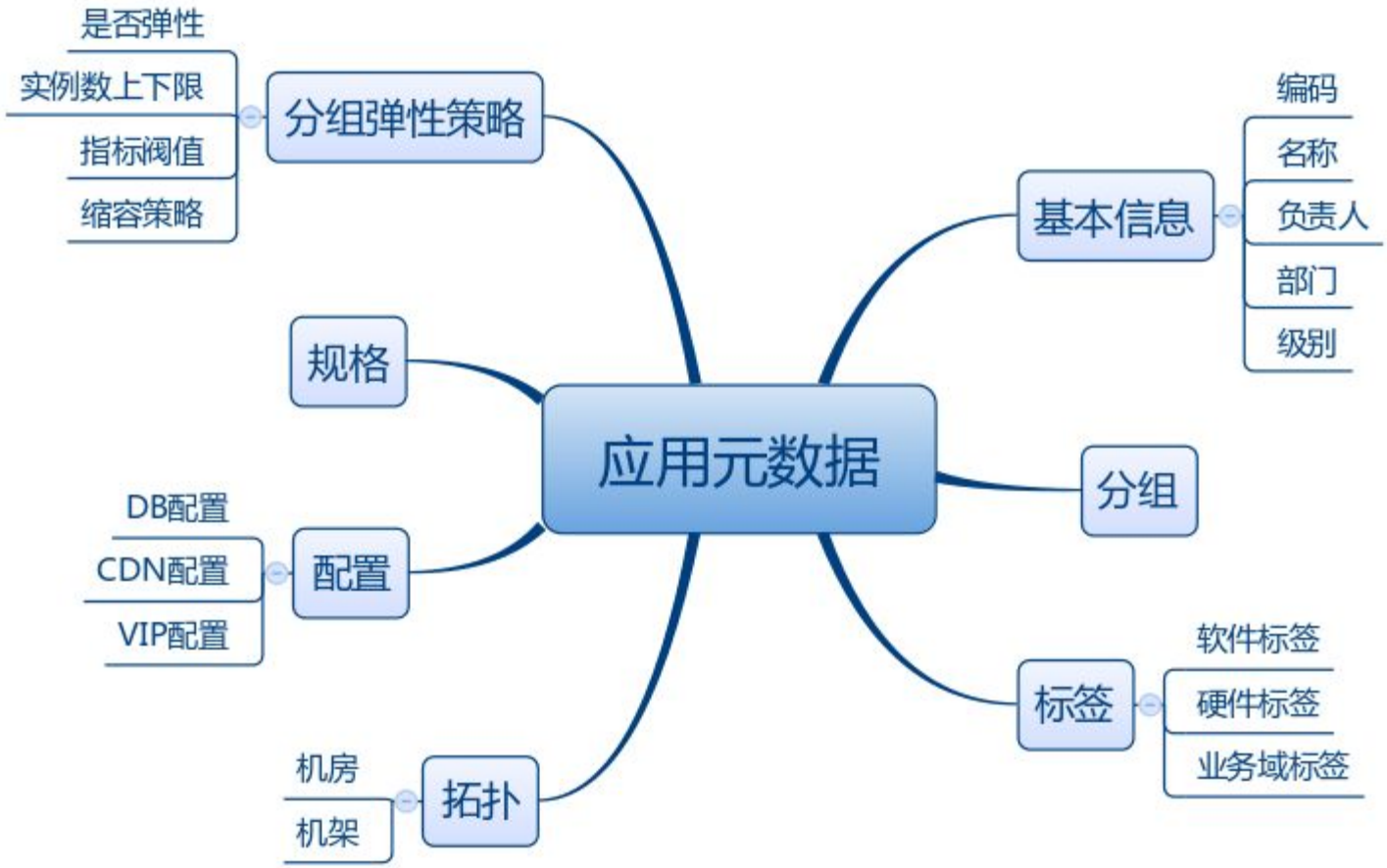
不同的环境有不同的调度流程，只有生产环境才开启弹性。



应用元数据

JD.COM 京东

一个应用可以有多个不同的分组，每个分组都可以部署到不同的机房，不同分组对应的程序配置或版本是不一样的。在弹性过程中，需要知道应用的元数据。分组可以覆盖应用的标签、拓扑、配置和规格信息。



基于元数据的资源分配算法

JD.COM 京东

- 根据应用拓扑选择合理的机房和机架；
- 根据业务域标签、硬件标签和应用级别选择合适的域；
- 根据软件标签选择合适的基础镜像；
- 根据规格（CPU，内存，磁盘）限制容器占用的资源。

弹性调度策略和算法

JD.COM 京东

弹性调度单元是应用分组在一个机房的实例。

■用户可以配置分组是否弹性，该分组实例在各个机房的上限和下限，扩容的指标最小阀值，缩容的指标最大阀值，每次缩容的数量或比率；

■只有正式环境开启弹性，上线日禁止弹性；

■根据该应用分组在指定机房的整体负载情况，采用2次移动平均算法来预测下一时刻在该机房的负载来进行弹性，预测数据时间范围为7天；

■集成第三方JSF扩展接口，由第三方根据应用性能指标提供预测的建议，实现弹性微服务。

议题

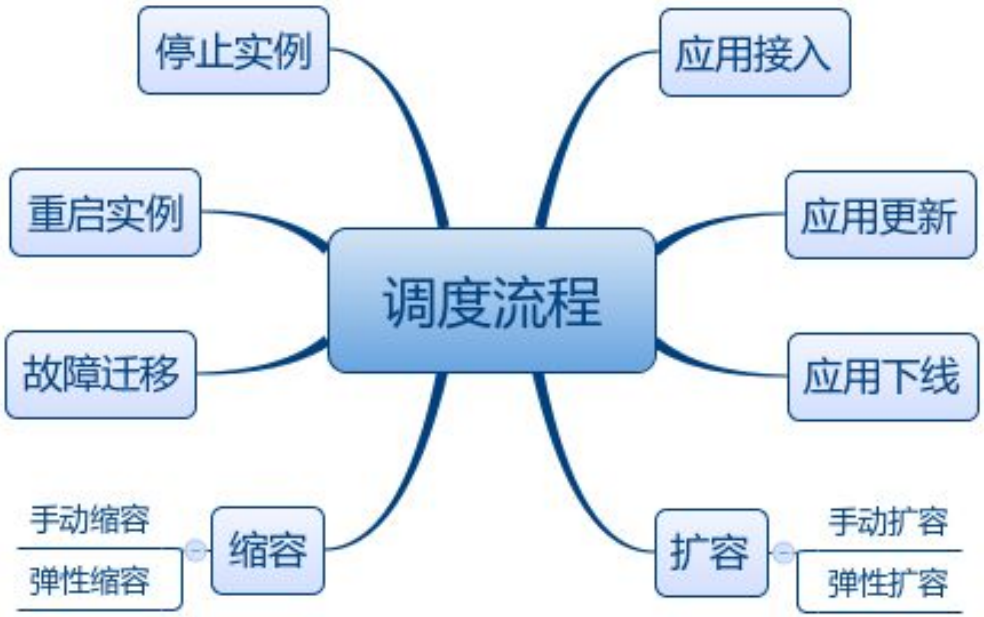
JD.COM 京东

1. 京东弹性计算之路
2. 京东弹性计算架构
3. 京东弹性调度策略及算法
4. **京东弹性调度流程**
5. 京东弹性计算应用场景

弹性调度流程

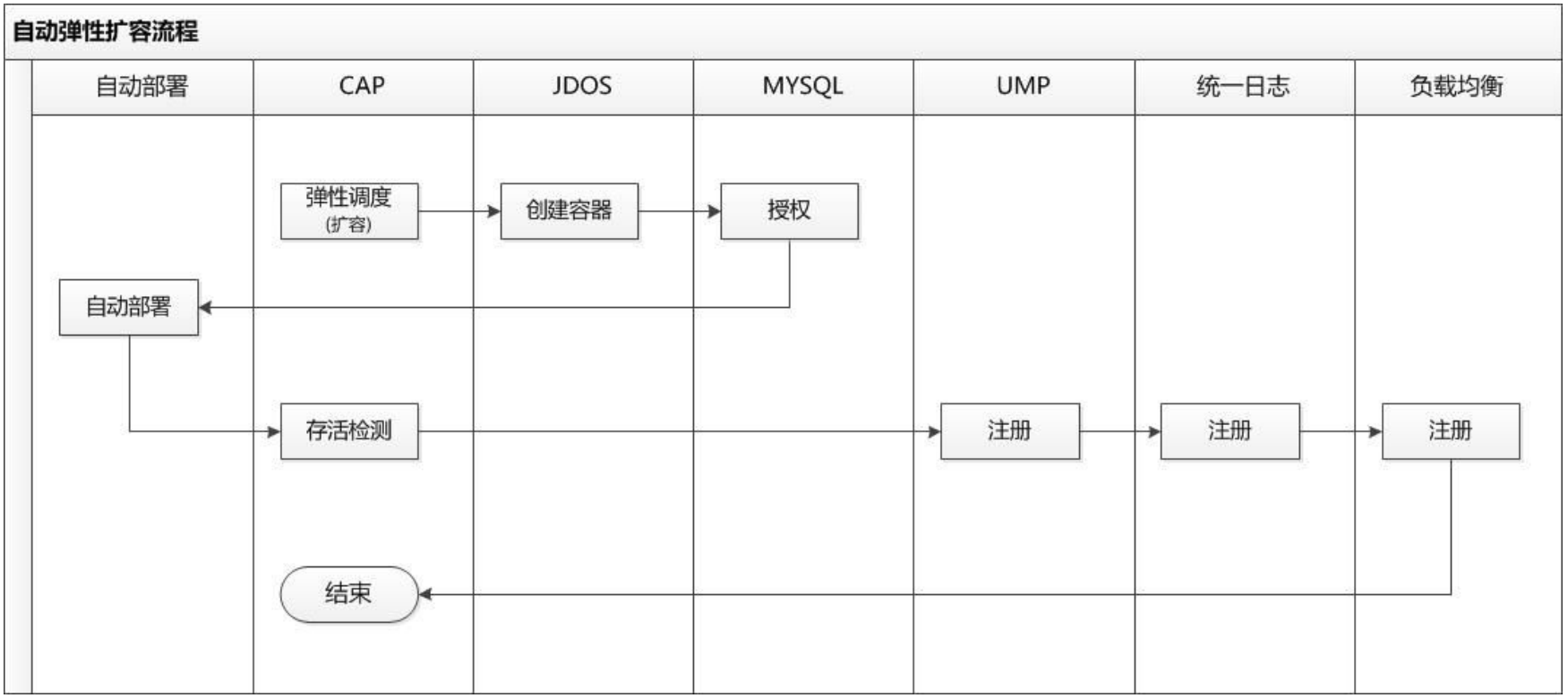
JD.COM 京东

- 涵盖了应用的生命周期：初次接入，日常的运维更新，下线。
- 支持手动和自动两种伸缩模式，应对不同的应用需求。秒杀类应用可以提前使用手动扩容，活动结束再缩容。



弹性扩容流程

JD.COM 京东

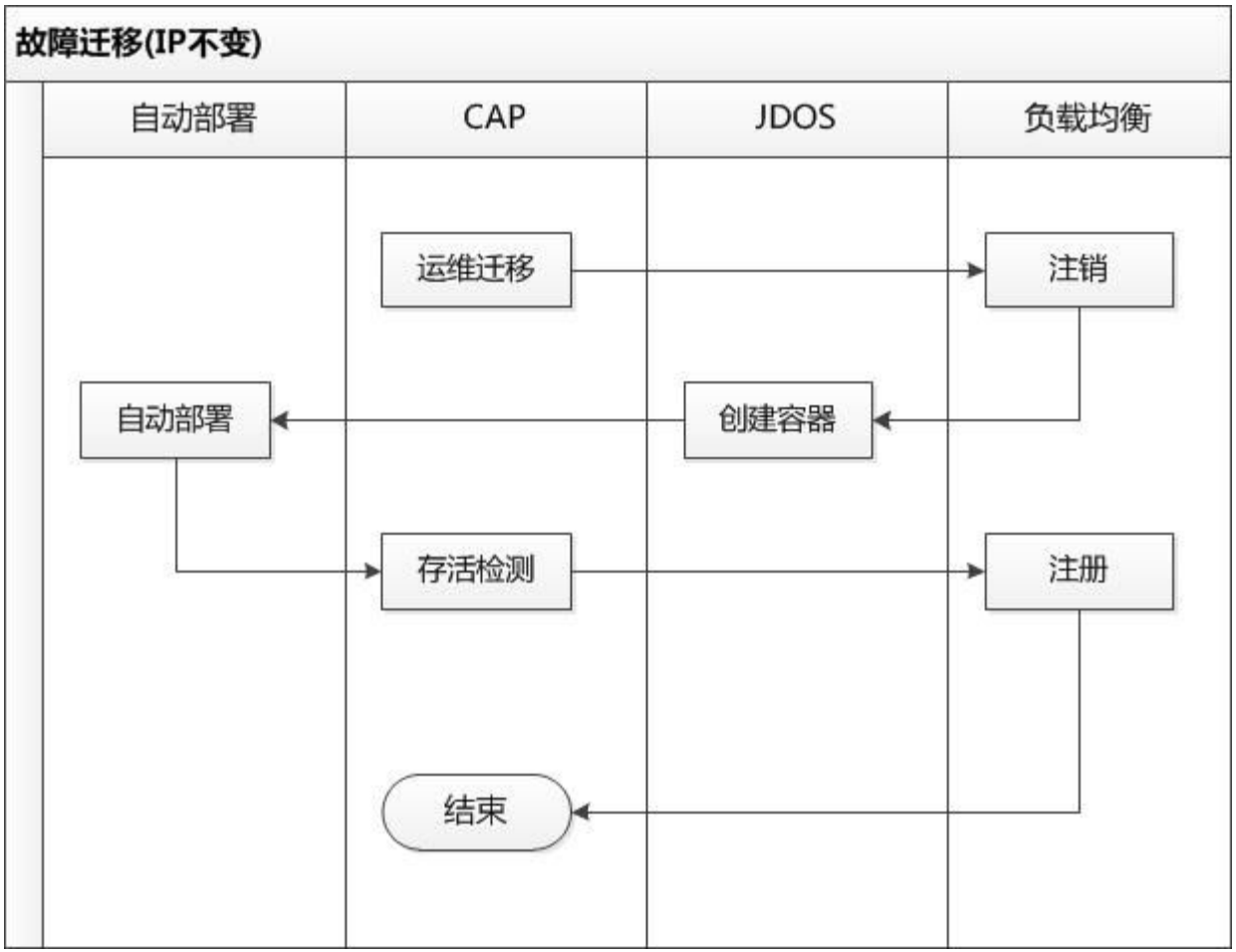


应用在启动之前可能需要数据库授权，启动之后需要挂载VIP，注册统一监控和统一日志。如何能自动发现应用的注册信息，采用了模版方式。应用先申请一个容器，手工注册这些信息，后续的扩容会以该容器为模版来进行自动注册

故障迁移流程

JD.COM 京东

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。



议题

JD.COM 京东

1. 京东弹性计算之路
2. 京东弹性计算架构
3. 京东弹性调度策略及算法
4. 京东弹性调度流程
5. **京东弹性计算应用场景**

弹性计算应用场景

JD.COM 京东

- 京东弹性计算经过了618的大流量考验，包括：图片展现80%流量，单品页50%流量，秒杀风控85%流量，虚拟风控50%流量，618作战指挥室大屏
- 目前接入了网站，交易，订单履约，配送，售后，无线，拍拍，O2O等等核心应用
- 廊坊新机房全部采用弹性云建设，将支持双11大流量



弹性计算实践经验

JD.COM 京东

- 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；

谢谢！

www.jd.com

JD.COM 京东