京东云原生跨域大数据平台落地实践

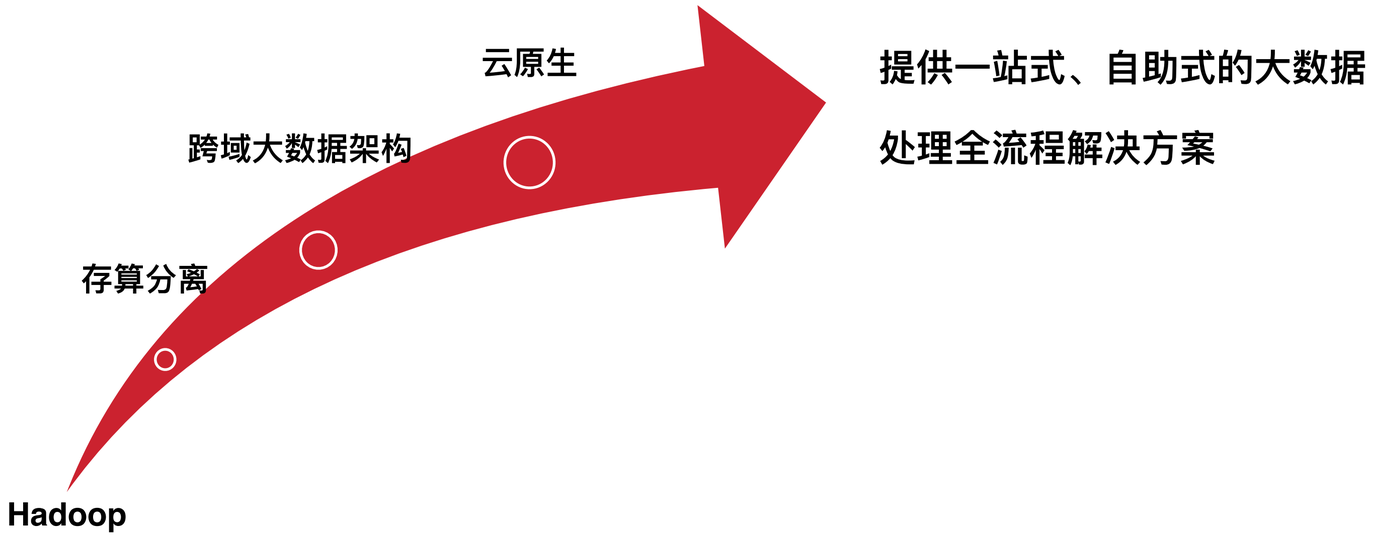
京东零售-集团数据计算平台部 / 吴维伟

目录

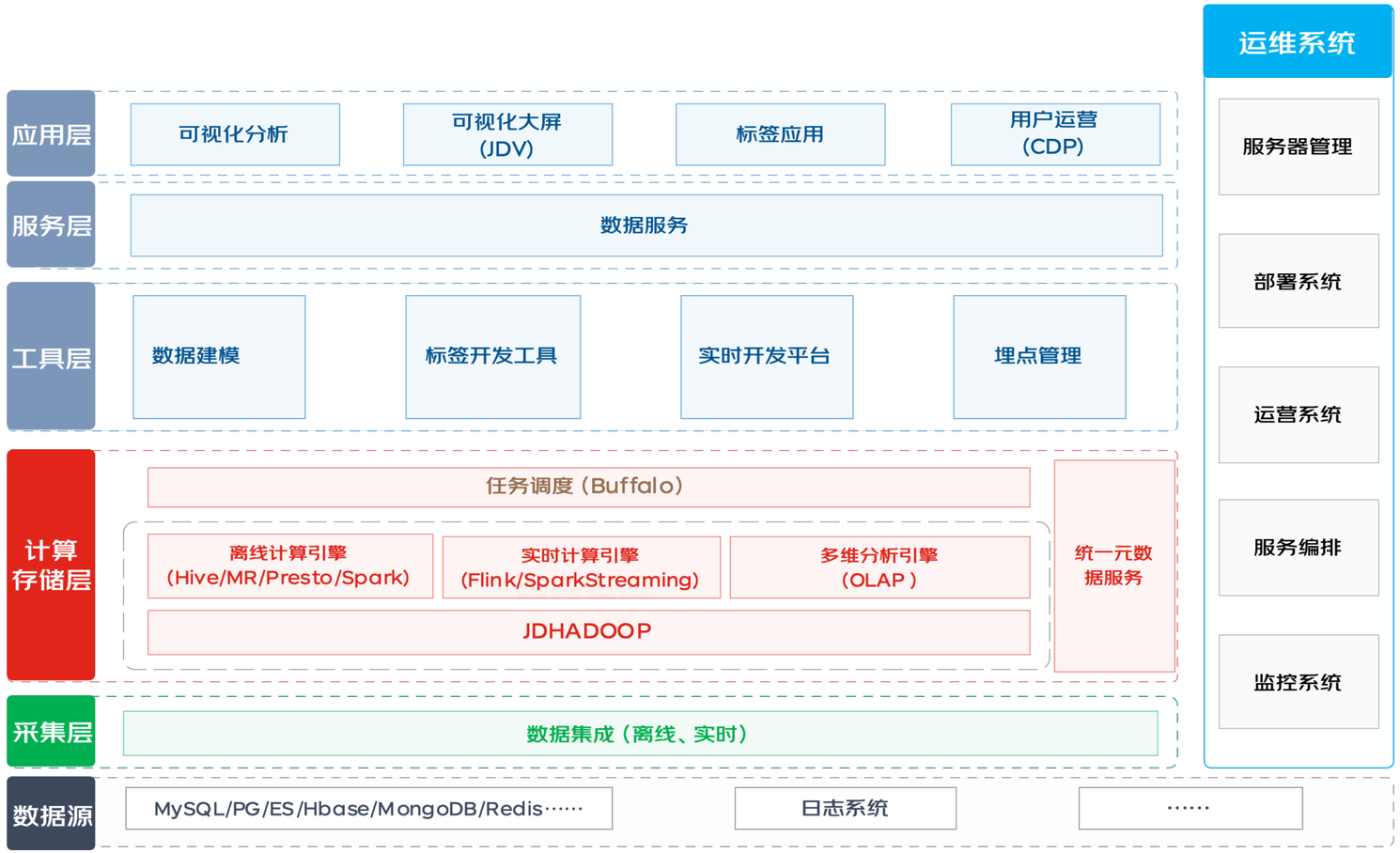
- 一. 京东大数据平台概况
- 二. 京东云原生大数据平台建设背景和挑战
- 三. 京东云原生大数据平台落地实践
  - 离在线混部
  - 跨域存储
- 四. 落地收益
- 五. 未来规划

一. 云原生大数据平台概况

京东大数据平台是京东大数据业务的基础服务平台，为京东大数据业务的实现提供一站式、自助式的大数据处理全流程解决方案。涵盖数据采集、存储、加工、分析、可视化、机器学习等专业化产品和服务，通过数据集中从而形成高效的数据开放，在保障数据安全的前提下，提供自助式的服务平台，大幅降低大数据消费门槛，帮助京东大数据业务快速落地，助力京东实践以数据为驱动的业务变革与发展。



一. 云原生大数据平台概况-平台架构



集群规模
数百万核

计算能力
日运行job数百万

存储能力
数 EB

二. 云原生大数据平台-建设背景

痛点

- 大促期间在线业务采购机器应对业务高峰
- 日常在线资源利用率低

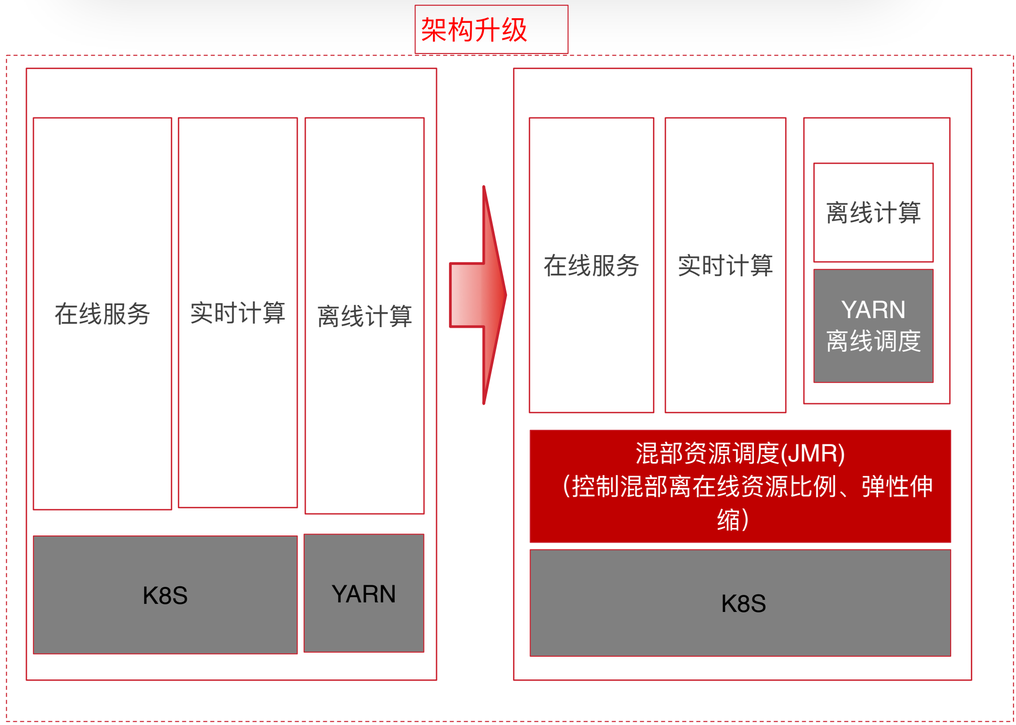
目标

- 大促期间在线业务 0 采购扩容
- 日常在线资源共享给离线使用

=

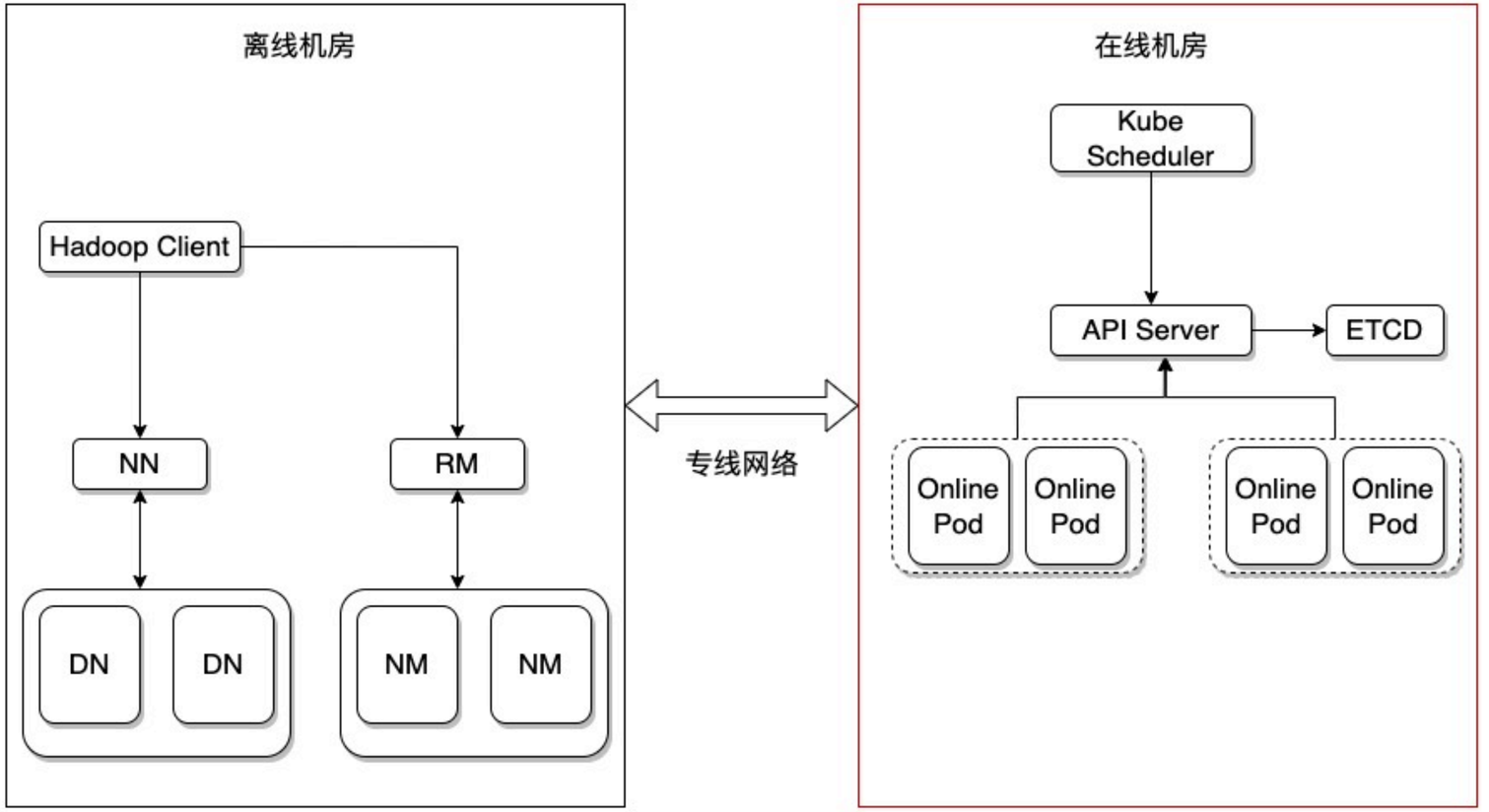
云原生弹性伸缩架构升级

资源弹性伸缩的最大价值是通过资源共享方式充分复用资源，在资源使用上的错峰填谷，以达到提升集群资源利用率的效果，实现资源价值最大化。



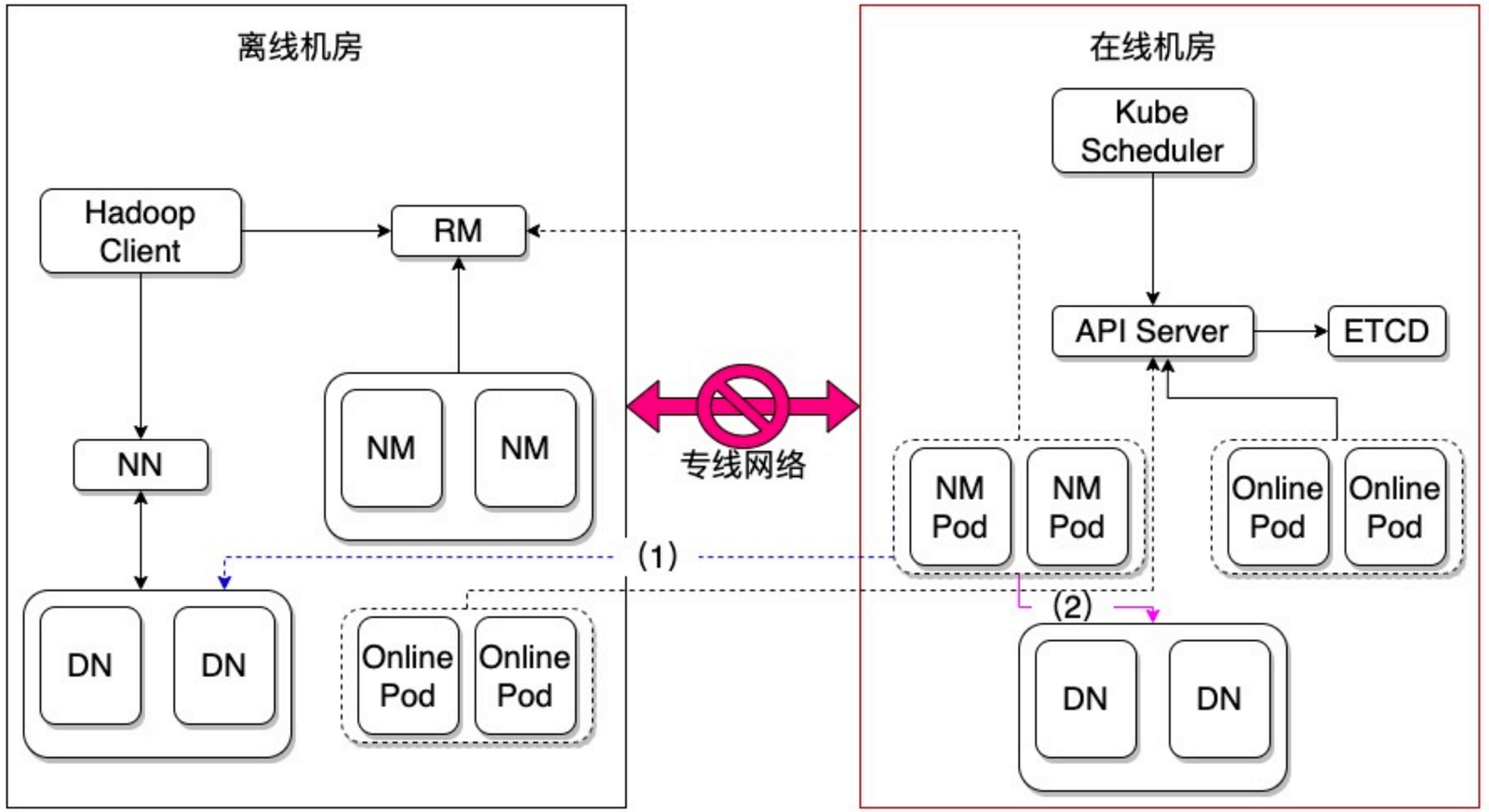
二. 云原生大数据平台-建设挑战

- 如何统一离线和在线的资源调度?
- 离线在线混合部署时，如何保证在线业务不受影响，离线业务基本稳定?



二. 云原生大数据平台-建设挑战

- 跨机房资源共享后，跨机房数据访问如何避免影响在线任务（网络隔离与流控）



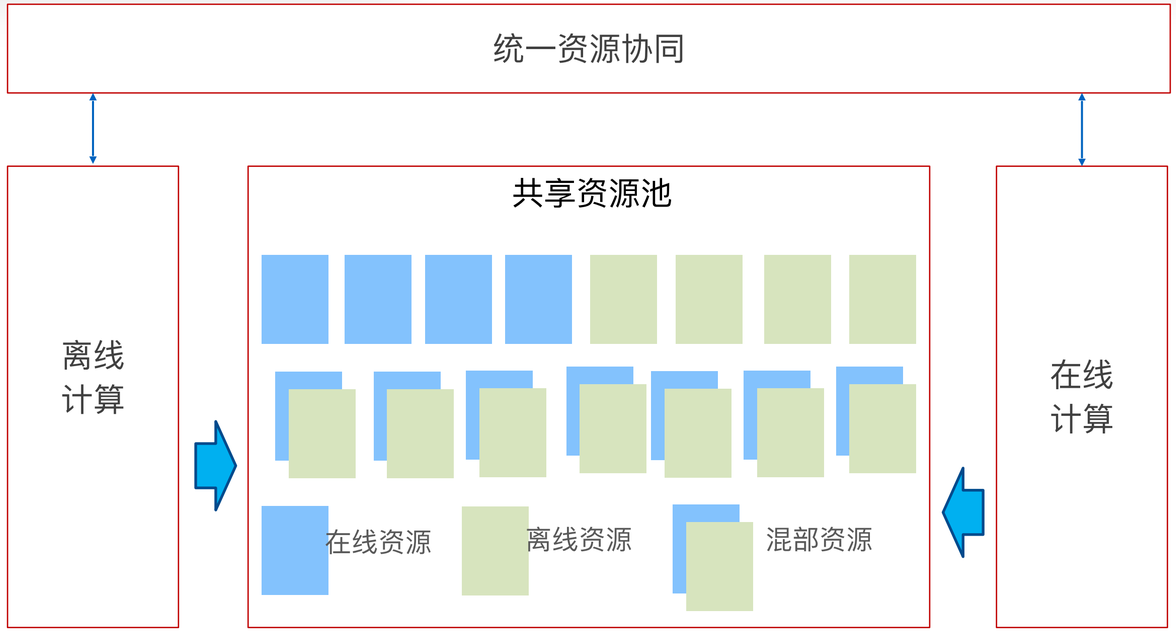
三. 落地实践 -（1）计算混部

资源池化

- 资源统一封装，屏蔽底层IaaS特性
- 统一资源调度，上层应用系统无感使用
- 按需调度，大促节点，离线仅需借出数小时资源

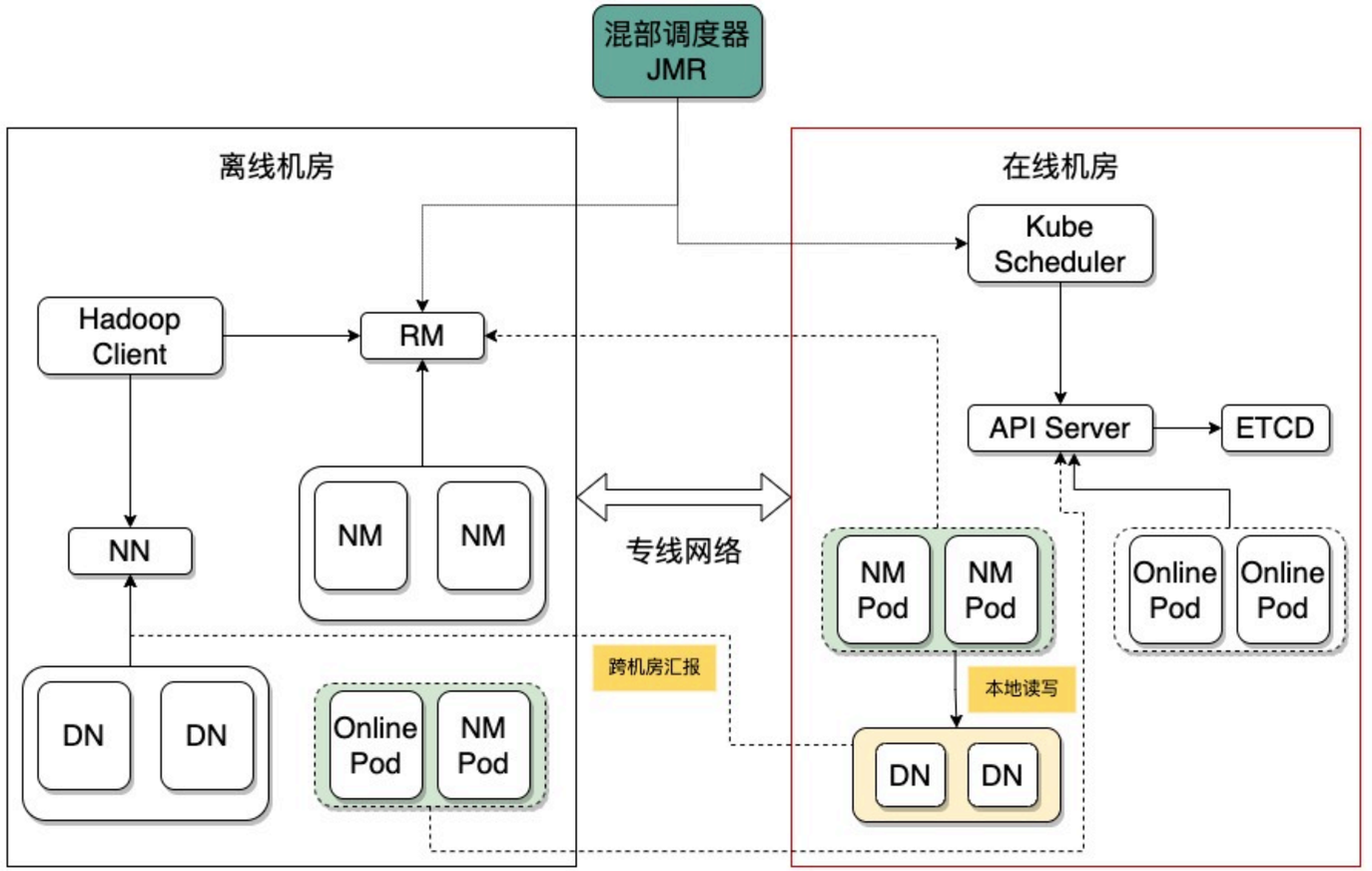
高效利用

- 在线应用和离线计算具有资源互补的特点，可通过统一资源调度提升资源复用率



三. 落地实践 -（1）混部架构

- K8S 统一资源管控
- JMR（混部资源管理）协调混部资源调度，结合单机弹性实现资源动态伸缩。
- 强资源隔离保障在线业务 TP99



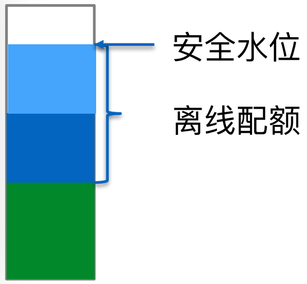
三. 落地实践 -（1）混部关键技术

统一资源管理

- K8S 统一管理资源
- 计算服务容器化改造
- 混部调度器（JMR）协调 K8S 资源分配和 NM 弹性伸缩

单机弹性

- 安全水位
- 离线最小最大配额（min，max），动态调整
- 定制化驱逐策略：容器类型、优先级、启动时间、资源容量



资源隔离

- 联合 K8S 团队实现 CPU 隔离、网络 QoS，保障在线业务 TP99
- 改造 HADOOP 底层，支持基于任务等级、流量类型等多种方式设置网络优先级

运维优化

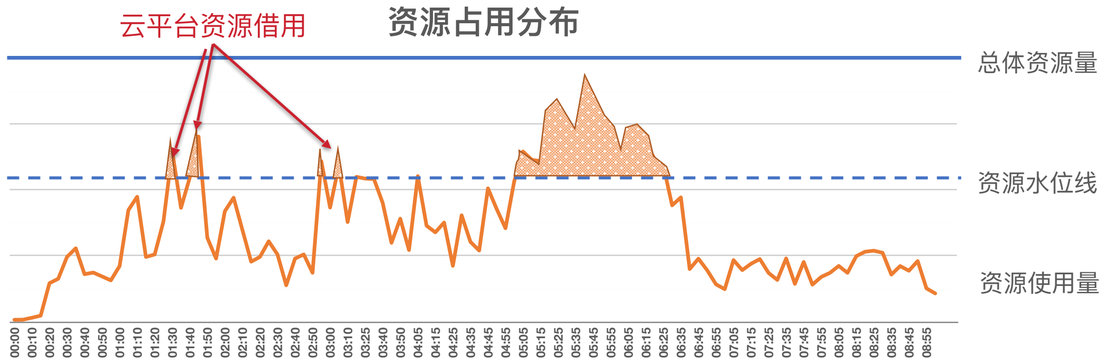
- YARN Operator 管理 NM pod 生命周期
- 基于 Token 方案实现 NM 节点注册验证

三.落地实践 -（1）混部资源动态规划

问题：资源占用分布不均衡，大部分离线资源长时间闲置

目标：利用弹性伸缩能力，峰值资源按需向云平台购买，减少离线计算常驻资源量

挑战：大规模、复杂作业链路，超百万任务，资源预测困难



基于作业分级，结合资源预测、数据血缘、作业性能诊断等能力，智能动态向云平台按需购买资源，降低离线机房常驻资源需求

技术内核

- 时序预测
- 任务诊断
- 数据血缘
- 动态规划

智能编排引擎

任务解析	日志解析	链路图谱
	运行时间	任务优先级
时序编排	资源预测	时长预测
	时序预测	……
资源调度	资源监测	读写监测
	任务托管	……

收益

- 利用云原生架构实现峰值资源按需购买，购买资源每日占用时长不超过 4 小时
- 资源水位线以上计算资源可以分批腾退，降低大数据平台机器成本

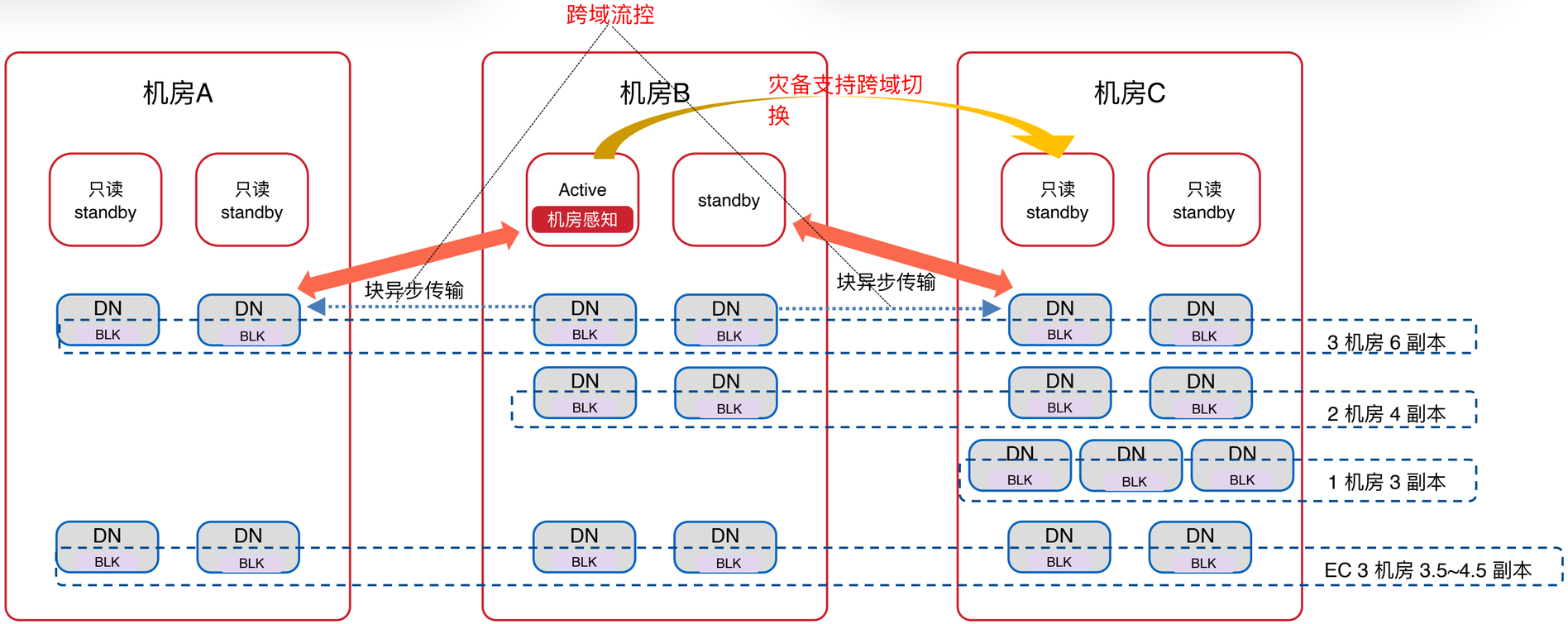
三. 落地实践 -（2）跨域存储

架构优势

- 跨机房读取变为本地读取，减少跨域流量
- 跨域生命周期实现只同步最新数据，历史数据自动删除
- 支持数据机房级容灾

架构改造

- 机架感知->机房感知
- 跨域容灾：灾备可读，支持跨域切换
- 跨域流控
- 跨域EC
- 低冗余EC（1.16副本）



三. 落地实践 -（2）跨域存储：机房感知和标签

- 这个DN属于哪个机房?
  - 拓扑管理： /region/cluster/rack

- 这个客户端属于哪个机房?

  - 机房感知：

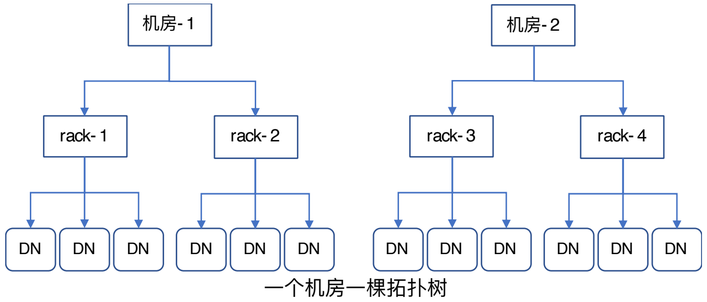
    - RPC 携带机房信息
    - 基于 IP 的机房查询

- 数据跨机房要怎么放?
  - 标识定义(支持副本及EC)：

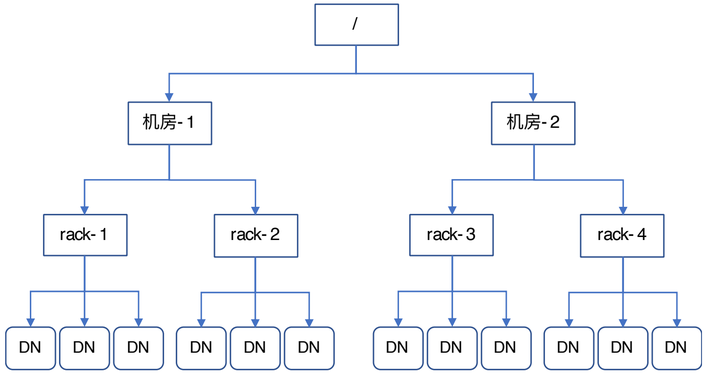
regionA:3:1, regionB:2:0,ttl:7200:regionA:2:1:MODIFY,ttl:7200:regionB:0:0:MODIFY

  - 元数据变更：

    - XATTR
    - 块属性标识



一个机房一棵拓扑树

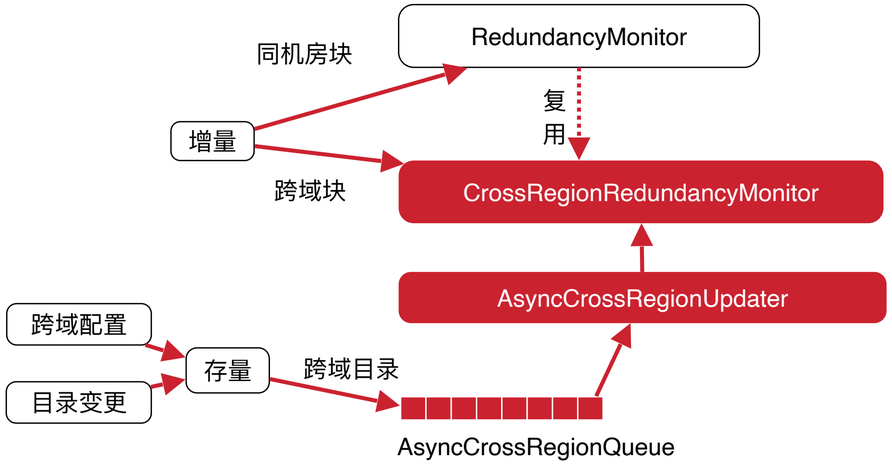


一棵大拓扑树

三. 京东云原生大数据平台-（2）跨域存储：数据分发及流控

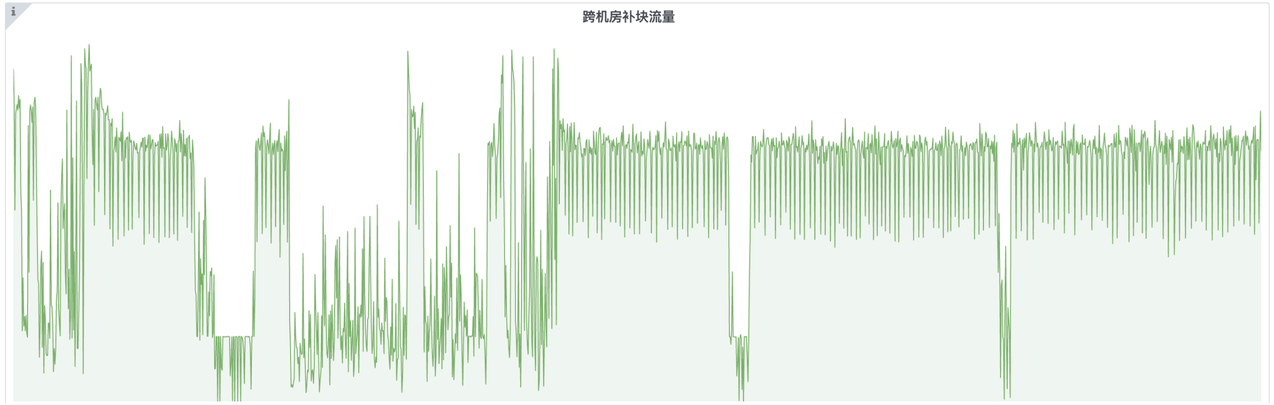
跨域补块

- 跨域补块独立处理，不影响原有同机房逻辑
- 异步跨域更新器，结合跨域标签属性，实现切换接续补块
- 支持高效的跨域数据共享

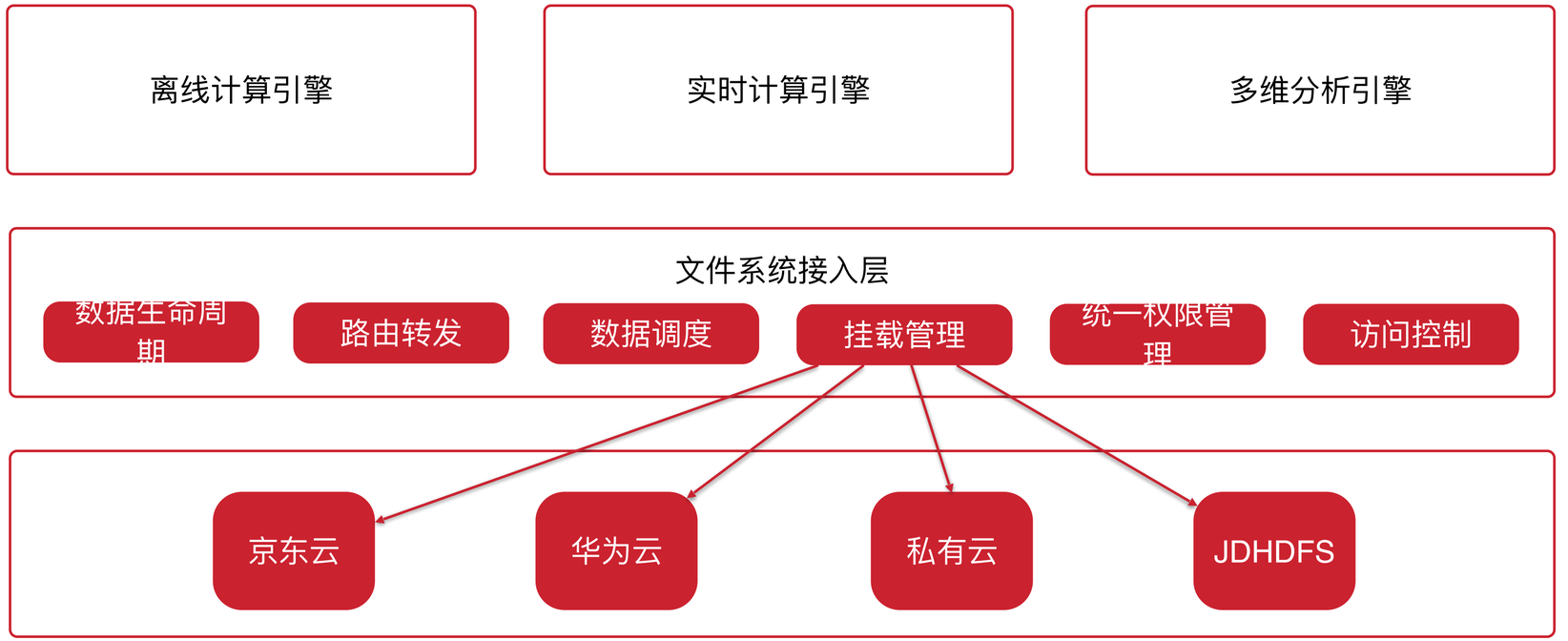


跨域流控

- 跨域补块流控
- 读写优先客户端同机房 DN
- 跨域读写流控
- balancer 机房内部均衡



三. 京东云原生大数据平台-（2）跨域存储：存储云原生

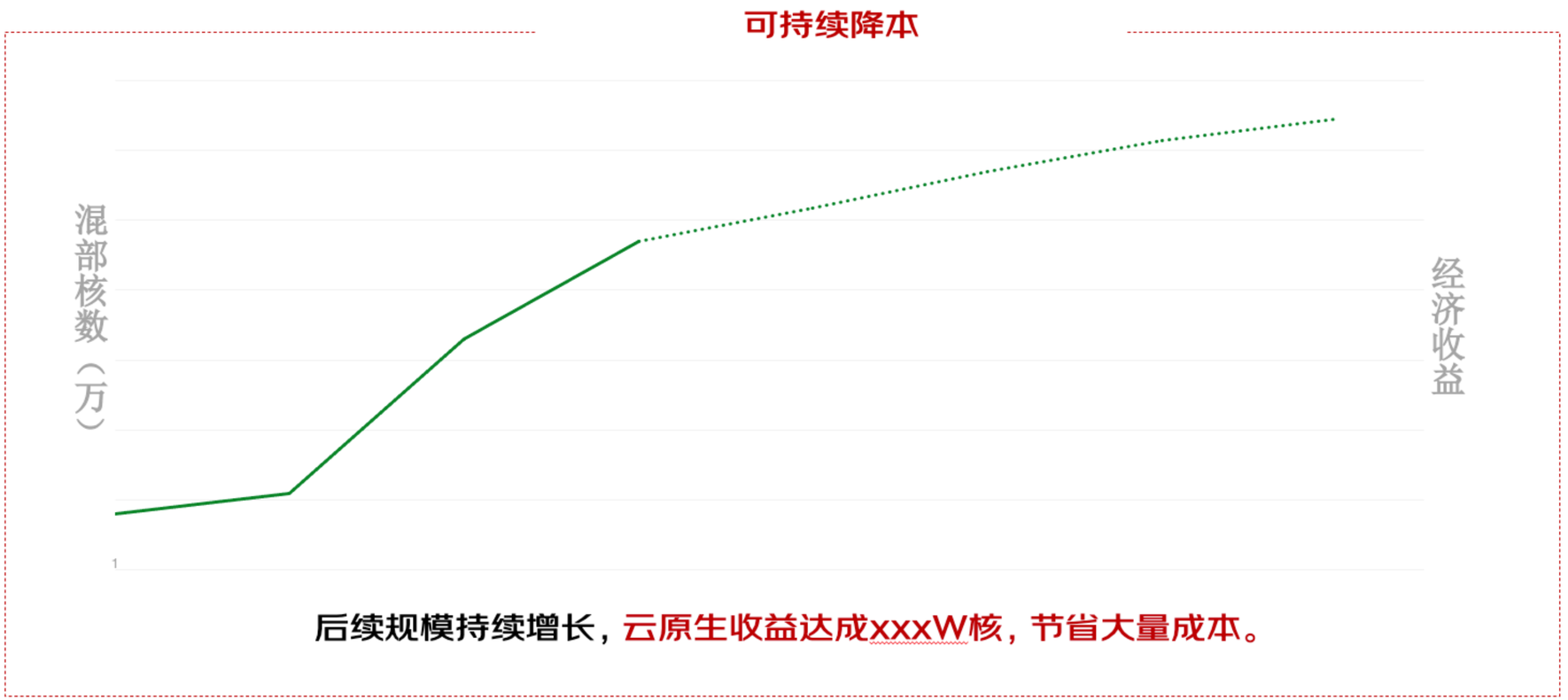


- 接入层实现通用需求，包括权限、访问控制、数据生命周期、数据调度等
- 接入层利用挂载能力实现弹性扩缩容
- 数据调度实现不同挂载存储的数据迁移

四. 落地收益

618及双11大促期间动态调拨离线平台数十万核支撑在线系统流量高峰，节省大量采购成本

日常期间，离线平台复用在线系统资源数十万核，利用率提升20%+，节省大量成本



五. 未来规划

阶段一

多云架构：弹性异构的云原生大数据

阶段二

湖仓一体：离线实时一体化数仓

阶段三

云数仓：达到低成本、灵活定义的架构优势

建立一个开发、建模、共享的数据生态，实现离线实时批流一体与湖仓一体的云数仓大数据平台

想一想，我该如何把这些技术应用在工作实践中？

THANKS