elastic 中文社区

Elastic 中文社区

深圳Meetup

2019/11/16

Saturday 13:00

合作伙伴

vivo 神鲸 Coca-Cola 可口可乐 Broadview www.broadview.com.cn

合作社区

segmentfault 创造属于开发者的时代 OSCHINA 掘金 珠三角技术沙龙 http://techparty.org

基于K8S的ES服务平台在头条的实践介绍

黄杨锋

2019.11.16

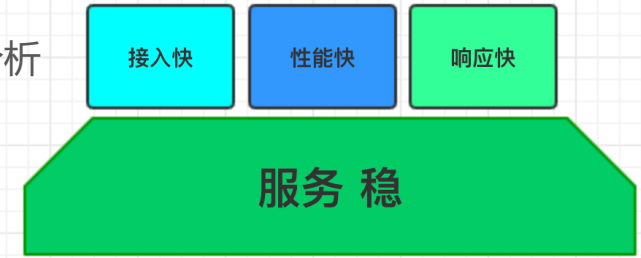
ByteDance字节跳动

目录

•定位与能力•

概述

- 定位：
  - 满足在线检索需求
- 目标：
  - 易接入：一键部署，白屏化操作
  - 可靠性：双机房容灾、主备同步、快照、隔离
  - 稳定性：切流、限流、降级、熔断、监控告警、日志收集分析
  - 易维护：扩缩容、数据导入、快速重建、自定义插件安装



愿景：

- 提供又快（接入快、性能快、响应快）又稳（服务稳定）的ES服务，为各个业务赋能。

Why

- **为什么要做ES平台?**
  - 现有的平台 性能、功能、稳定性不够
  - 解放业务，让业务专注业务
  - 满足在线检索需求，低延时

- **为什么要基于K8S来做ES平台?**
  - 隔离：业务隔离、资源隔离、影响隔离
  - 运维：快速部署、扩/缩容、升级
  - 资源消耗较VM低

已具备的能力

基础建设

- 具备监控告警功能
- 具备隔离特性
- 具有管理平台化能力
- 支持日志收集分析
- 支持按量计费
- 自动化部署
- 水平扩缩容

功能强化

- 支持GDPR
- 支持认证授权
- 支持索引重建
- 具备索引数据同步功能

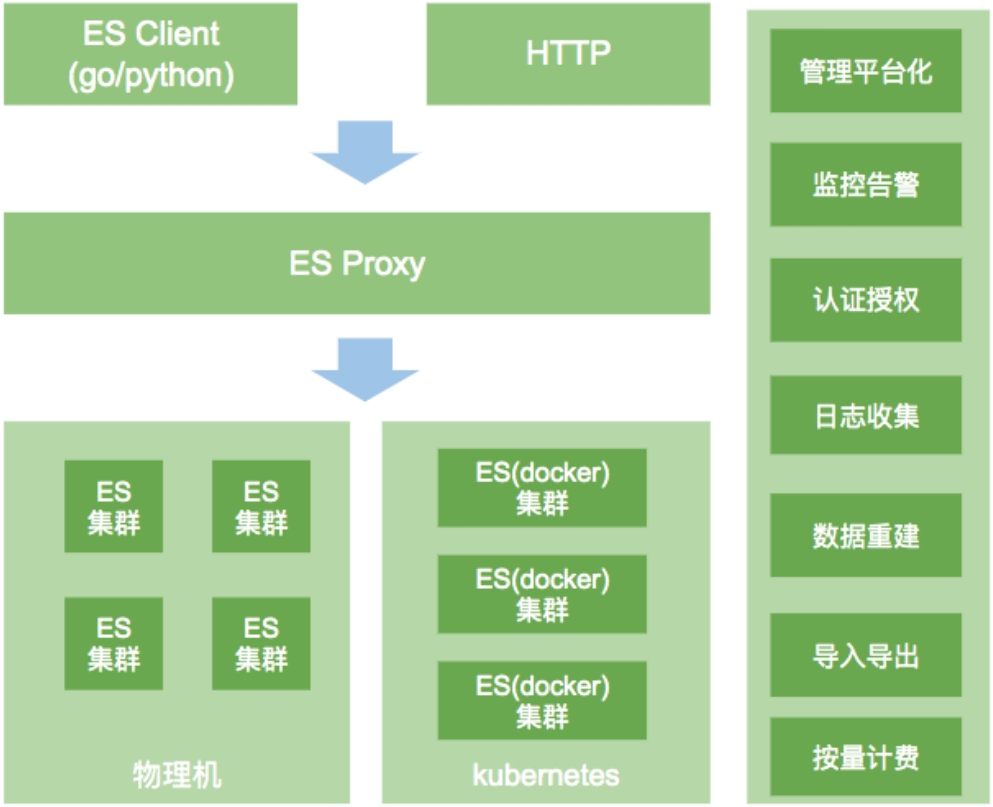
稳定性提升

- 支持削峰限流、切流、机房调度功能
- 支持按业务隔离部署
- 具备快照及快速恢复能力
- 具备跨机房容灾能力

•功能及模块介绍•

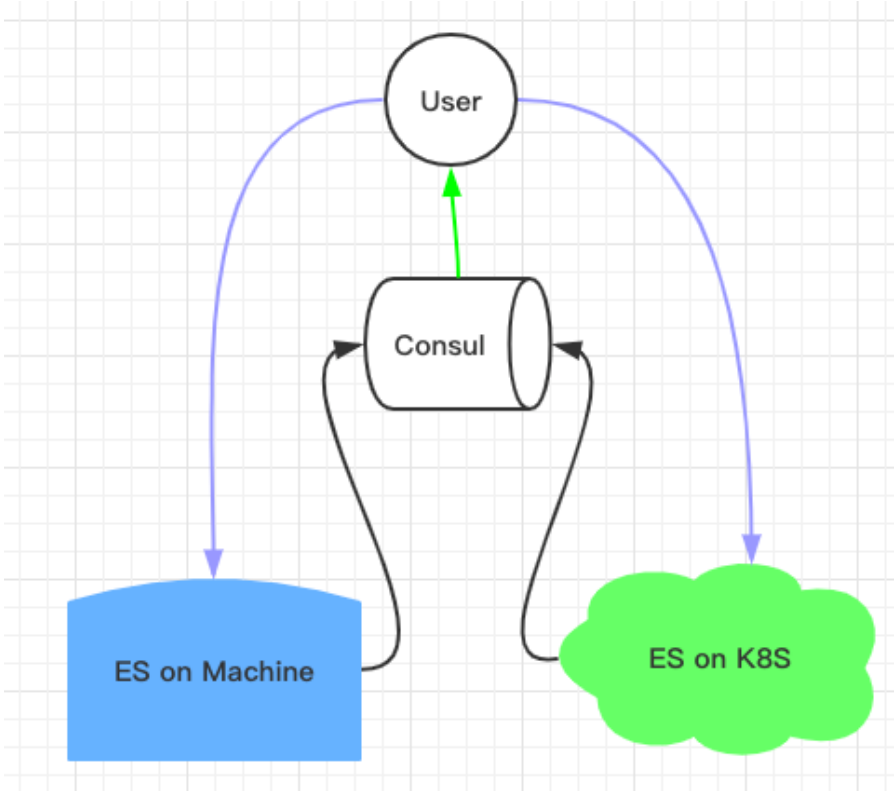
整体架构

- 调用侧:
  - 原生http api
  - go版client
- Proxy:
  - 机房调度、限流、降级等
- ES集群:
  - 物理机/容器部署
  - 双机房容灾，跨机房同步
  - 安全认证（GDPR）
- 配套系统:
  - 管理平台：集群\索引变更工单
  - 监控告警：打通公司服务治理平台
  - 日志收集：ES日志收集分析
  - 数据重建：基于快照备份&重建
  - 数据导入：流式写入&分词
  - 按量计费：按量付费/代持



ES集群的部署方式

- 物理机部署
  - 适用于集群规模较大、稳定性要求极高的场景
  - 完全独占的情形，不建议混部署
- 容器化部署：
  - 业务、资源、影响隔离
  - 快速部署
  - 水平扩/缩容
- 使用：
  - 用户通过consul 获得ES集群的ip 和port ， 对如何部署不感知



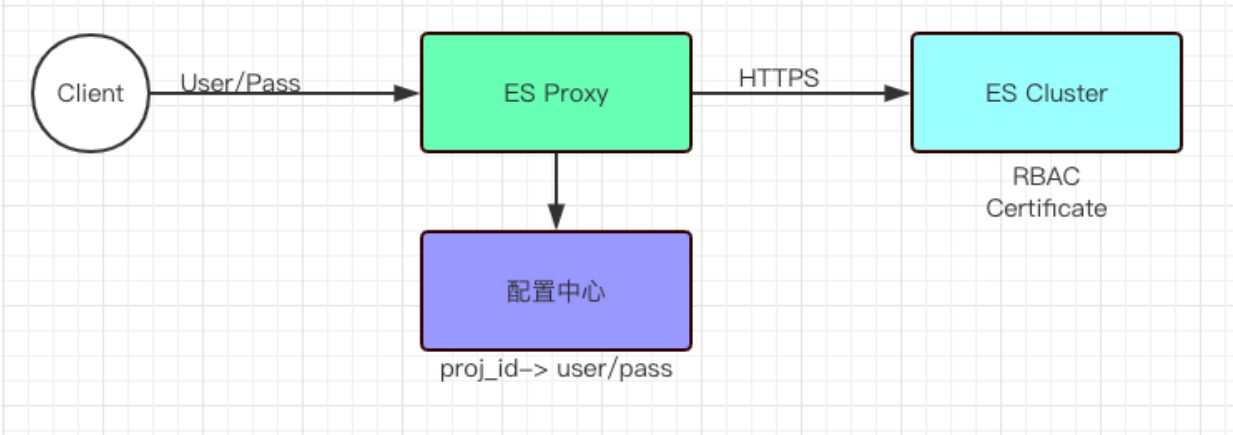
ES Proxy&Client

- Proxy: 流量接入层
  - API识别/信息注入(索引名/API/读写)
  - 与ES原生API保持一致
  - 具备限流、 降级
  - 机房调度：自动均衡/指定比例
  - 路由：按机房/索引
- Client: 基于HTTP & Consul访问服务

认证授权

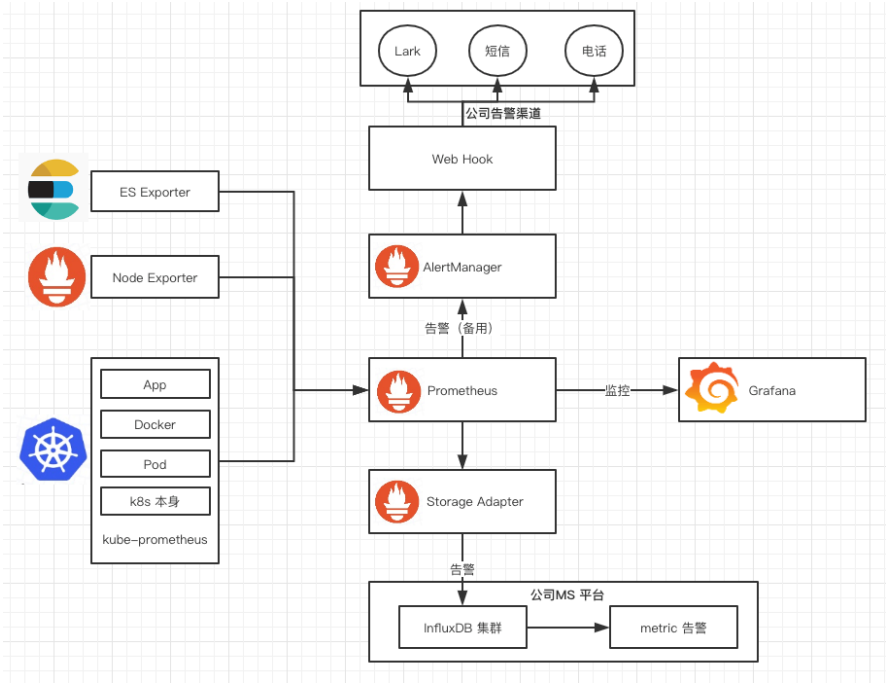
- Proxy层面校验
  - 读/写两种权限
  - 传输不加密
  - 安全性较弱，但无性能消耗
  - 默认开启

- ES层面校验：
  - 基于角色的访问权限控制(RBAC)
  - 控制粒度细，可精确到API层级
  - 传输要加密
  - 内部节点要认证
  - 安全性较强，但有性能消耗
  - 默认没开启



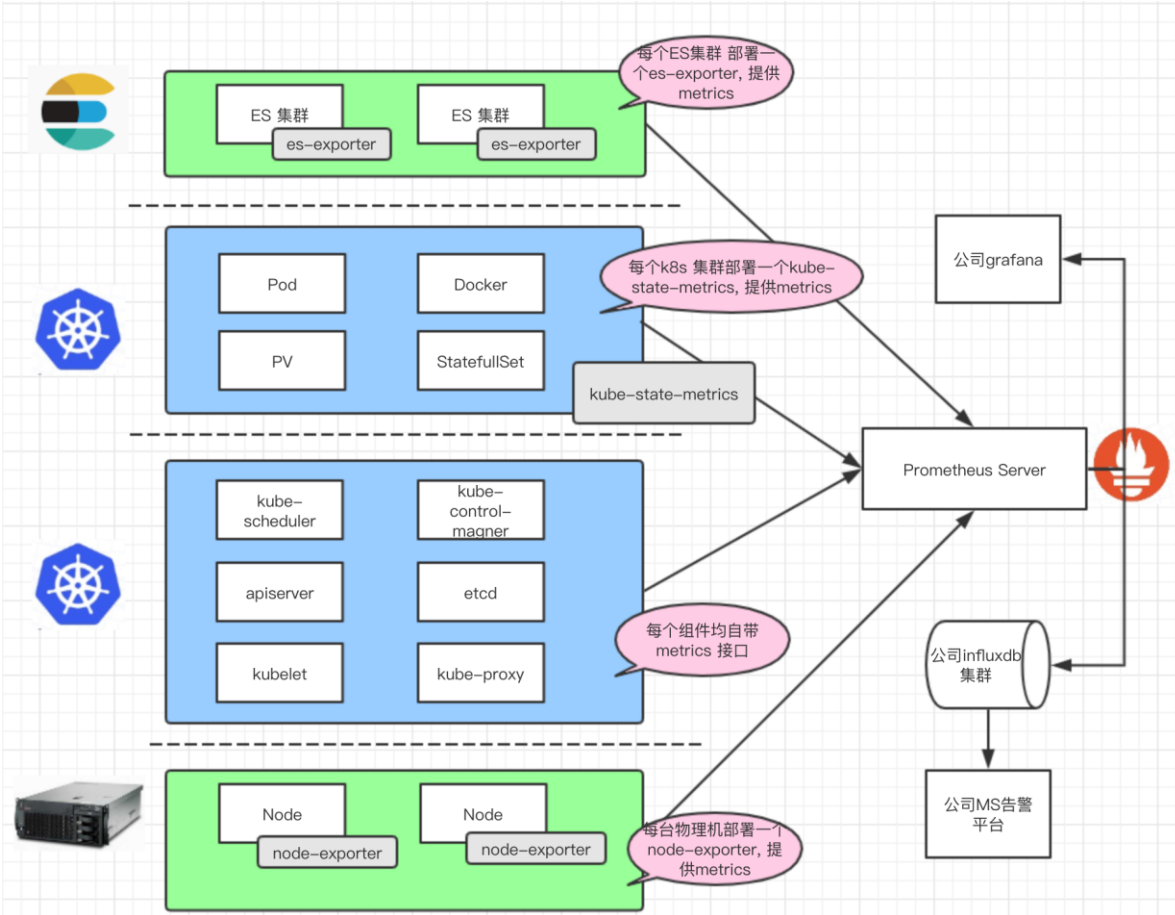
监控告警——物理机

- 监控：
  - 采用prometheus + XX Exporter + grafana
  - 涵盖metric生成、采集、存储、转存、分析、展示 等全链路功能
  - 支持物理机、ES本身及业务的监控
- 告警：
  - 将数据转存到InfluxDB集群，接入公司告警平台。
  - 实现告警自动注入。



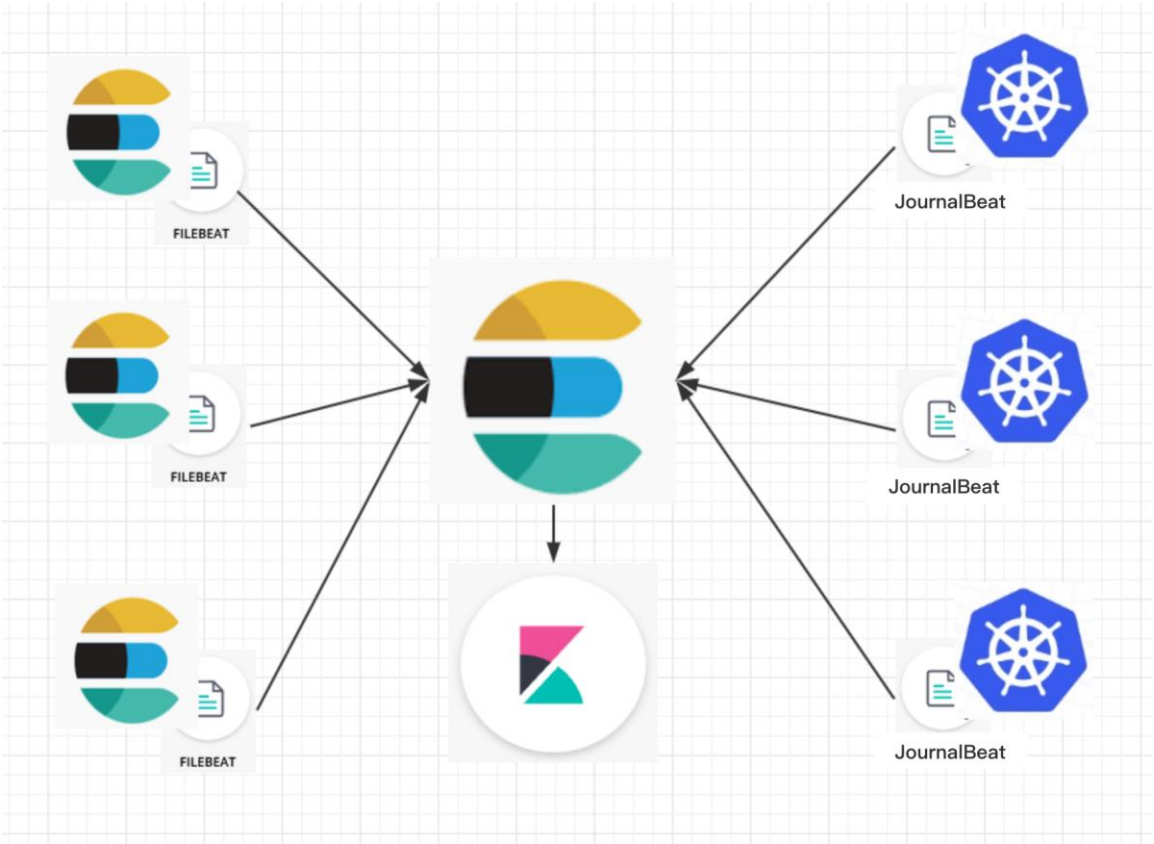
监控告警——容器

- 全栈式监控
  - 物理机metric收集
  - K8S组件metric收集
  - pod/pv/sts 等metric收集
  - ES metrics收集
  - 各个业务的 metrics 收集
  - prometheus自身的监控
- 告警
  - 组件存活告警
  - ES集群状态告警
  - prometheus自身存活/无数据告警



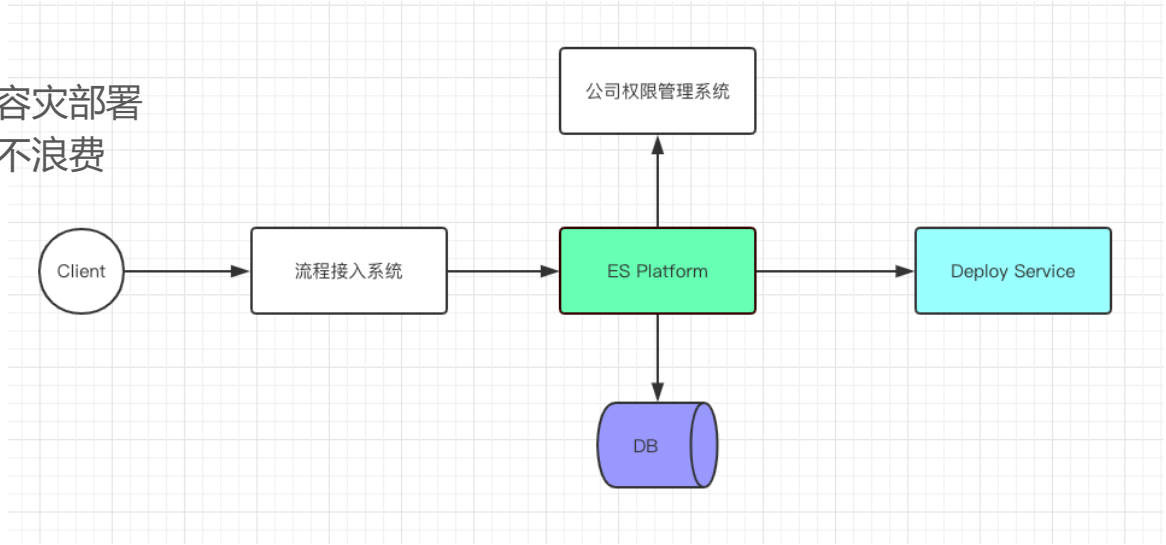
日志收集分析

- ES日志收集：
  - 包括slow log, gc log 等
  - 通过FileBeat 收集
- K8S日志收集：
  - 包括各个组件的日志，如kubelet, kube-apiserver等
  - 通过JournelBeat收集
- 日志分析：
  - 按业务、按天建索引
  - 日志数据清洗
  - 关键信息提取

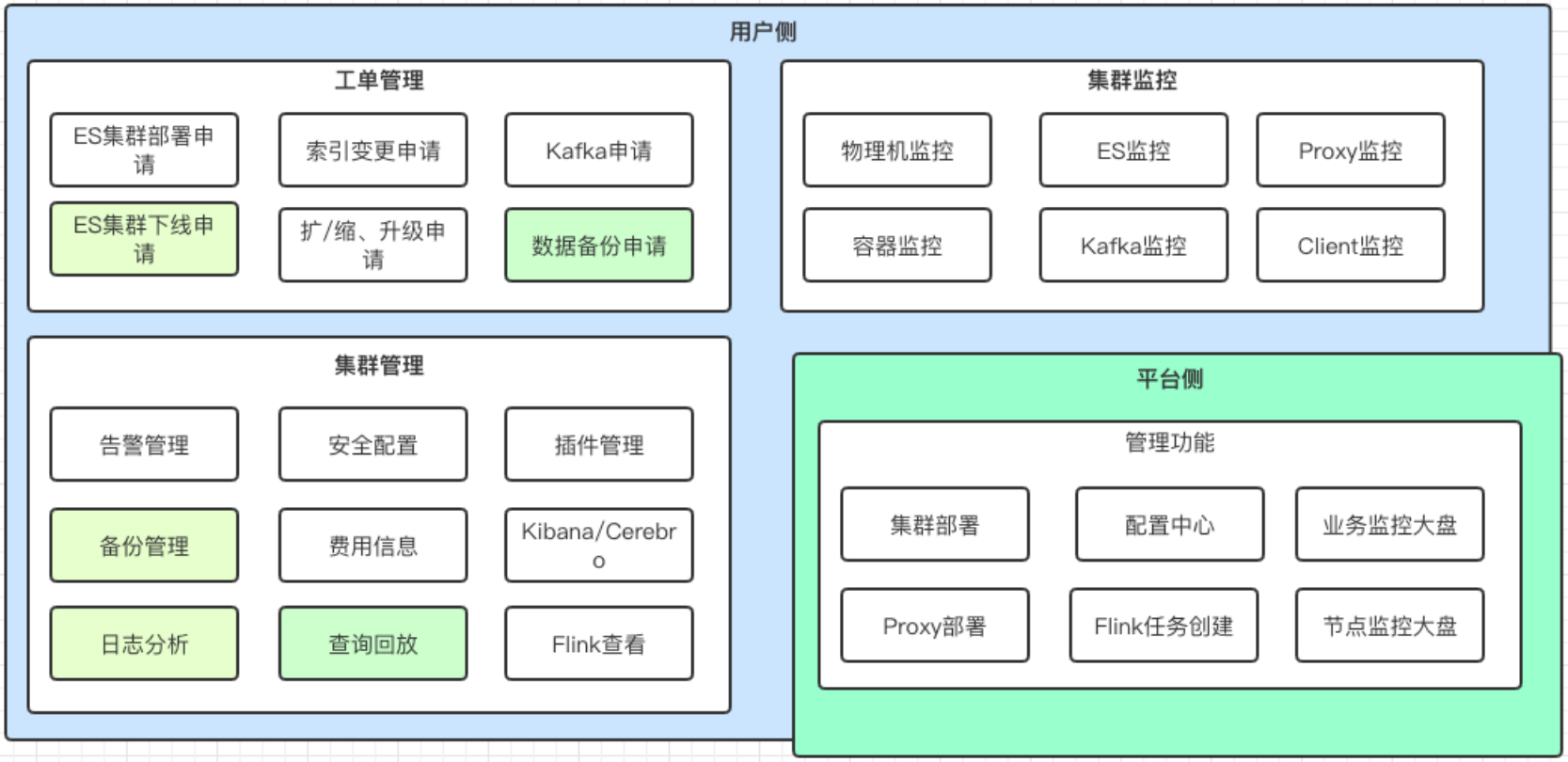


管理平台

- 平台特色：
  - 操作白屏化,低门槛
  - 记录操作日志，安全回滚
  - 点击鼠标即可完成多机房容灾部署
  - 多种实例规格可选，资源不浪费
  - 自定义插件安装
  - 接入权限校验
  - OnCall/智能客服
  - 文档站丰富(wiki)



平台化功能列表

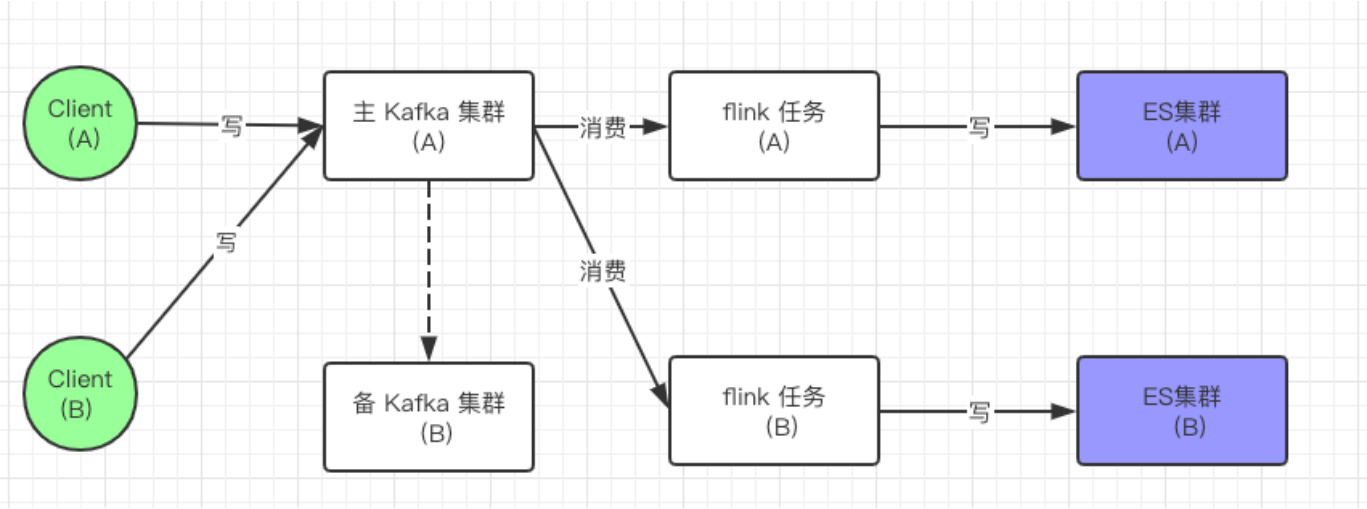


•重点与难点•

双机房容灾——双写（方案一）

特点：

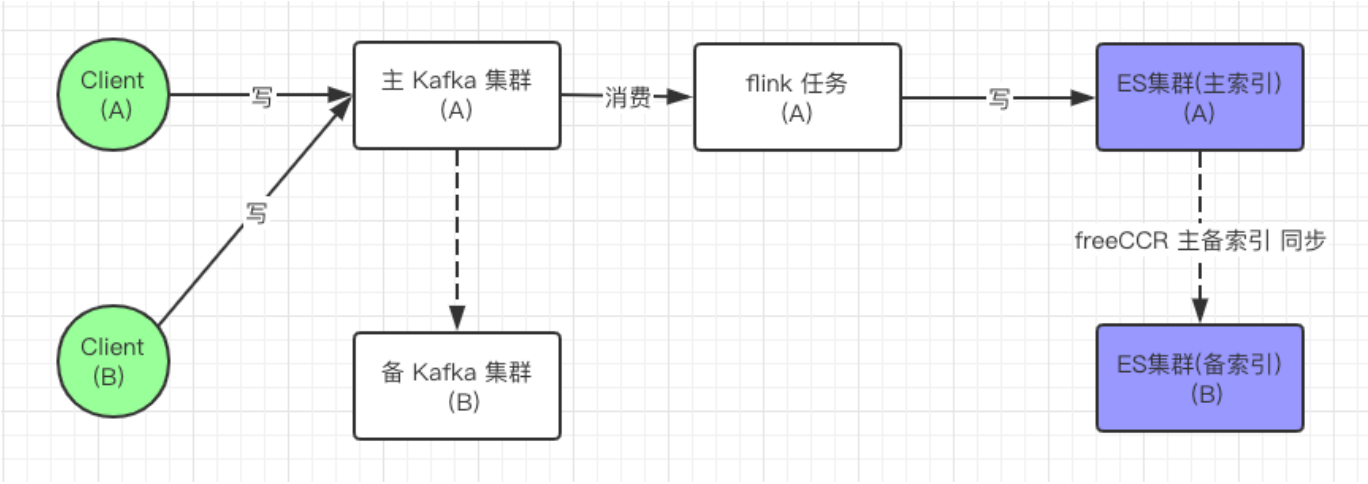
- 消峰
- 故障期间能继续读写
- 自动补写



双机房容灾——主备索引（方案二）

特点：

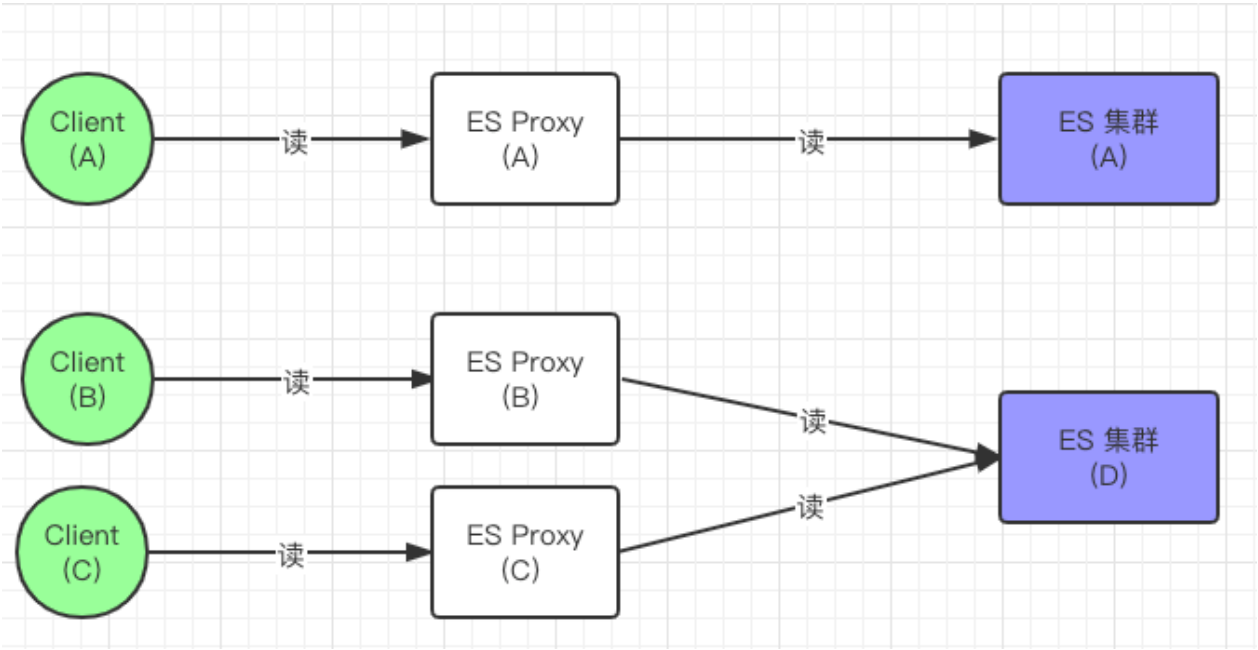
- 实时性
- 一主多备
- 对存量数据备份
- 异构mapping



双机房容灾——读流量调度

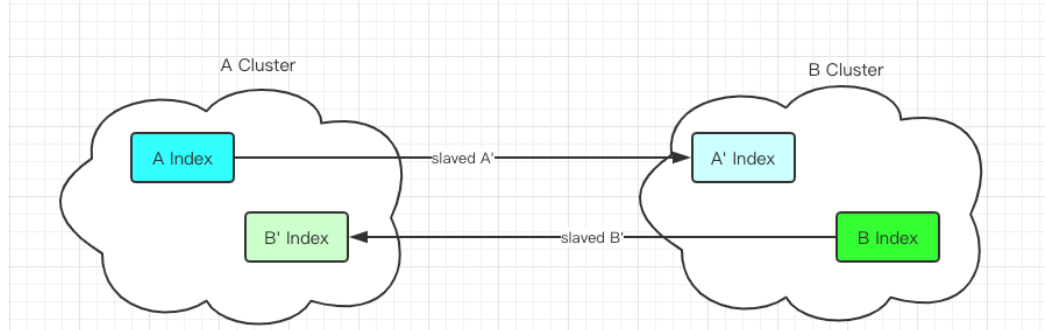
特点：

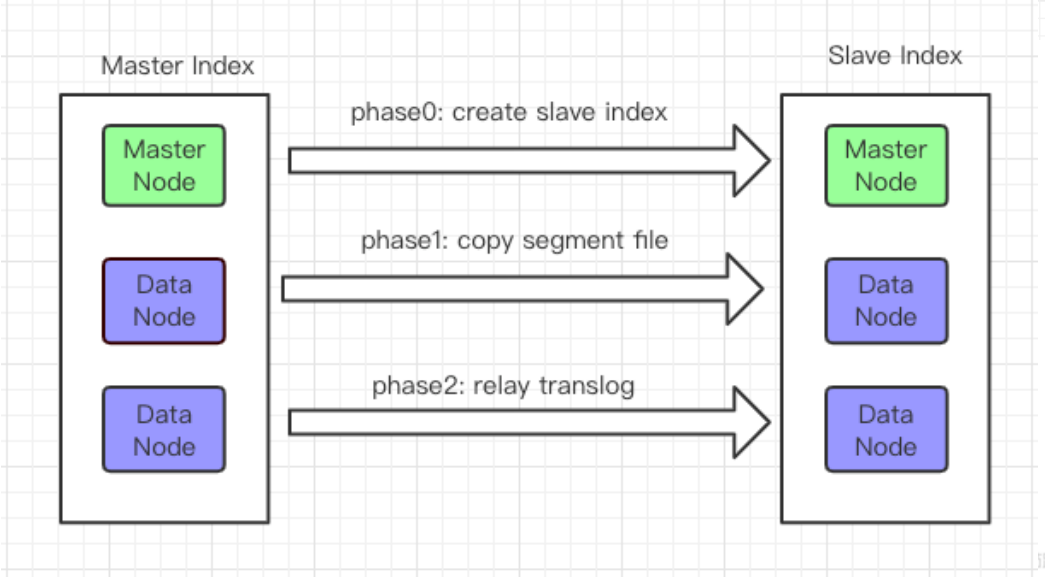
- proxy层调度
- 业务无感知
- index粒度切流
- 限流
- 降级
- 熔断



freeCCR 需求

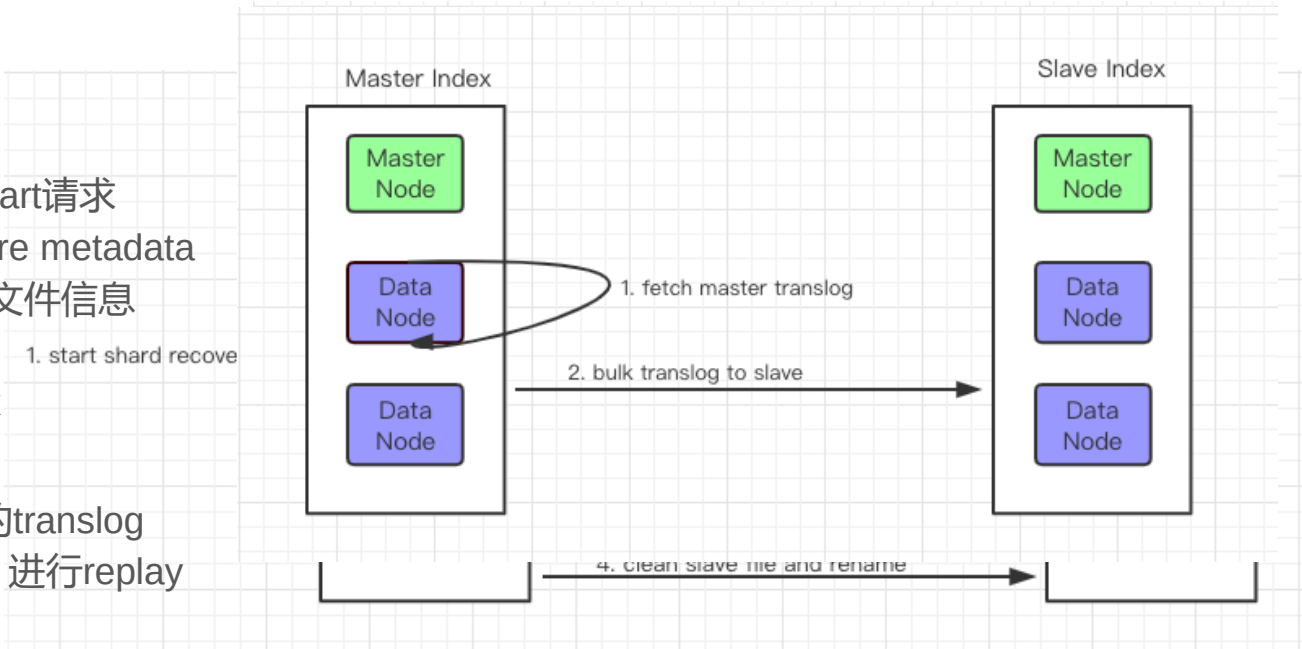
- 需求:
  - 多备分摊读压力
  - 支持同构/异构
  - 容灾需求
- 特点:
  - 索引维度的主备
  - 不支持主备切换
  - 采用推送方式
  - 延迟10s以内(压测时)





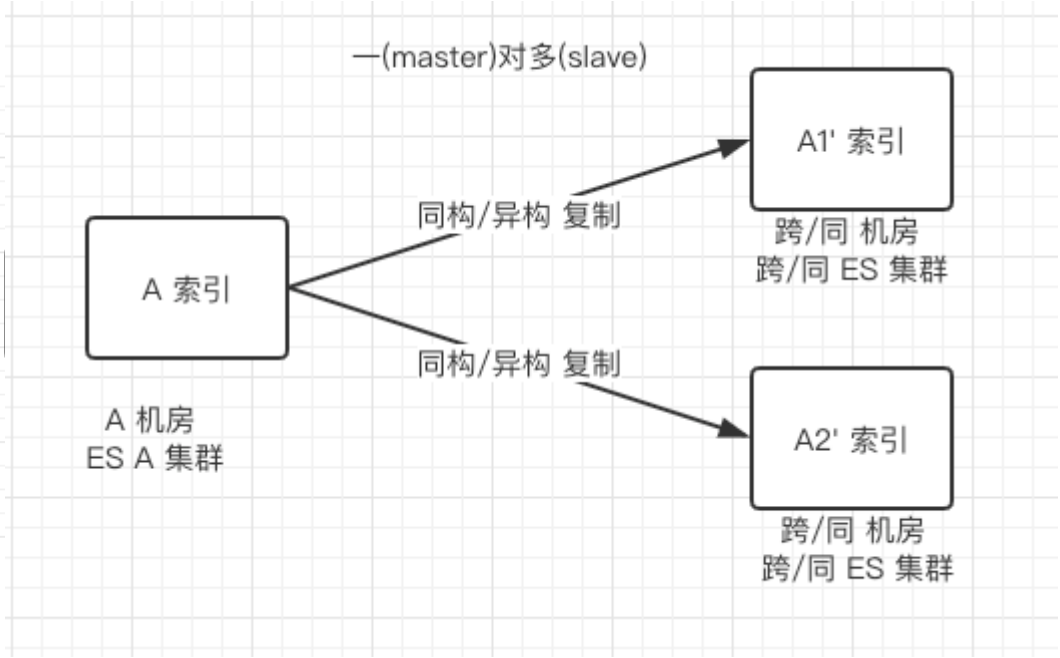
freeCCR 实现

- create slave index
  - mapping/setting
- copy segments file
  - master node发出start请求
  - 获得各slave 的 store metadata
  - 对比，发送缺失的文件信息
  - 发送文件
  - reopen slave index
- replay translog
  - 获得master index的translog
  - 发送到slave index 进行replay



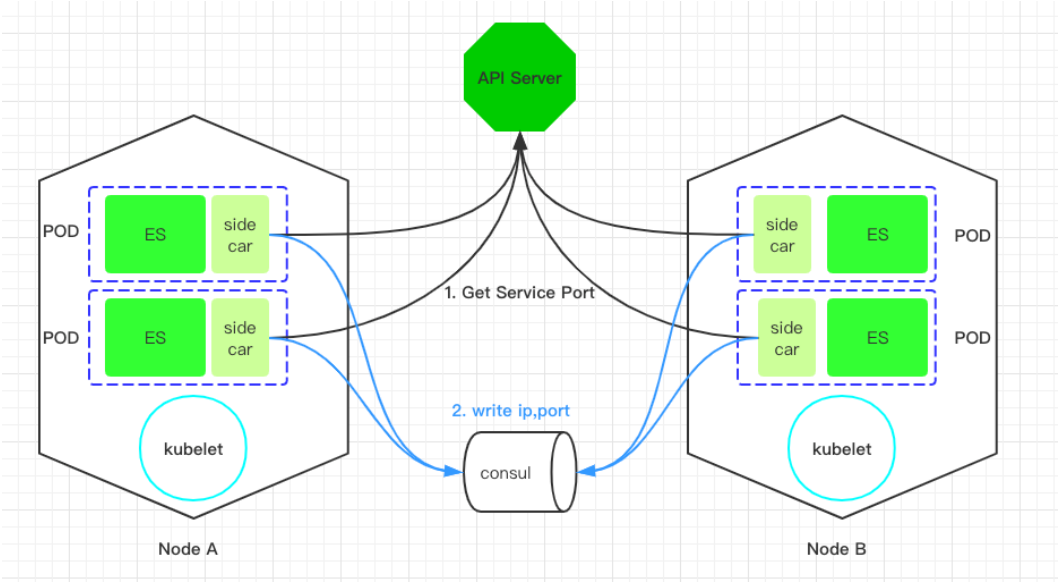
freeCCR 应用场景

- 主要场景：
  - 跨机房跨集群：应对容灾
  - 同机房跨集群：分摊读压力
  - 同机房同集群：异构复制，用于特定场景（分词效果、打分模型）
- 衍生场景：
  - 一主多备
  - 级连复制
  - 同构/异构混合



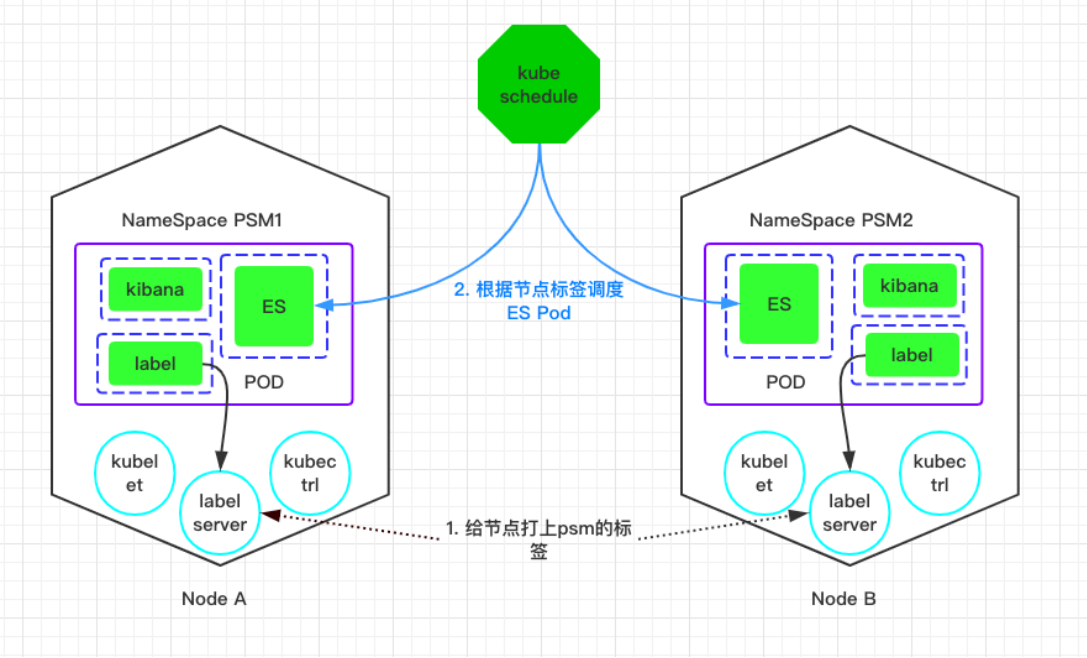
容器化部署（一）

- 服务发现：
  - 通过sidecar容器从api server获得svc port
  - 将物理机IP 传入sidecar 容器
  - 再将物理机ip和svc port 注册到consul
- 持久化存储：
  - ceph fs / block （性能不足、不适合大量小文件）
  - local pv （暂时不太成熟）
  - host path (调度时不会考虑磁盘信息)



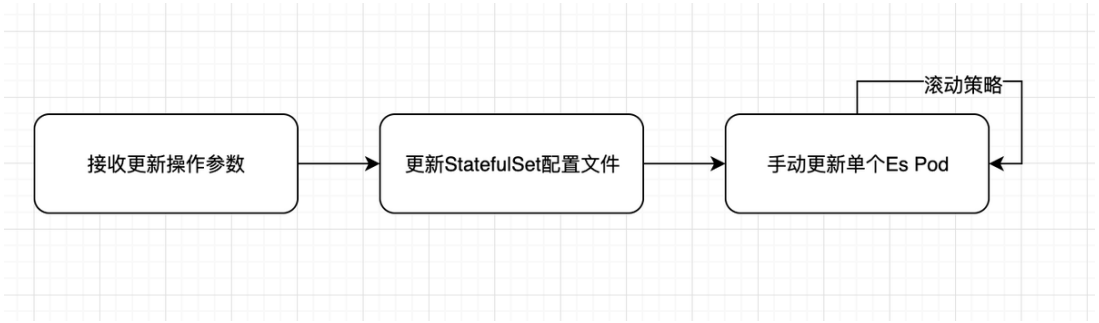
容器化部署（二）—— 漂移控制

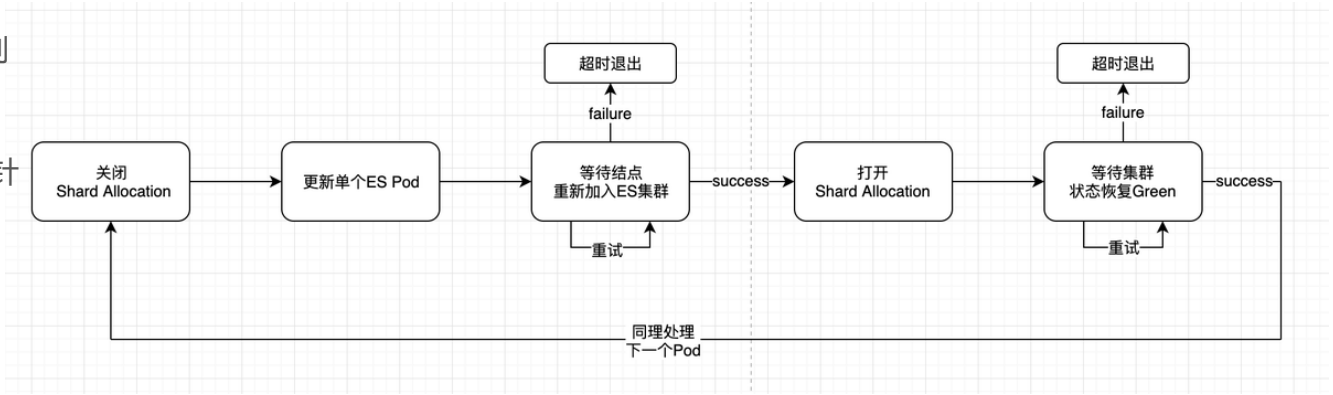
- 需求场景：
  - pod优先漂回原物理机
- 实现要点：
  - 每个物理机起一个label server 服务，用于给节点打上/取消psm的标签
  - 每个NS启动一个label docker给label server发指令（定期更新标签，prestop删除标签）
  - 为ES pod设置节点亲和性(Node affinity)



容器化部署（三）—— 完全可控的滚动更新

- 需求场景：
  - 水平扩容 (scale)
  - 垂直扩容 (set resource)
  - 更改镜像 (set image)
  - 更改环境变量 (set env)
- 实现要点：
  - 设置yml的更新策略为onDelete
  - 逐个删单个pod，待其重新创建
  - 待ES green之后再删另外一个pod。仅通过readness 探针是不够的。





•未来规划•

未来规划

- **短期规划：**
  - 查询回放
  - 定制化rank
  - 细粒度trace
  - 细粒度metric
  - 对用户透明的reindex
  - 基于成本的优化(CBO)
  - 结合业务深度定制

- **中长期规划**
  - 全内存查询
  - 实时查询
  - 借鉴vespa 进行取长
  - 机器学习建索引（Jeff Dean)

WE ARE HIRING

THANKS

ByteDance字节跳动

关于

Elastic 中文社区 http://elasticsearch.cn

Elastic Meetup 是由 Elastic 中文社区定期举办的线下交流活动，主要围绕 Elastic 的开源产品（Elasticsearch、Logstash、Kibana 和 Beats ）及 Elastic Stack 周边技术，探讨在搜索、数据实时分析、日志分析、安全等领域的实践与应用。

欢迎加入 Elastic 中文社区，**参与分享交流** 或 **赞助社区活动！**

深圳联络人：杨振涛
微信：nodexy
邮箱：nodexy@qq.com

本次活动回顾及现场照片在“vivo互联网技术”公众号发布，欢迎关注浏览。

微信扫码关注