知乎
www.zhihu.com

# 知乎基于 Kubernetes 的 Kafka平台探索和实践

知乎 白瑜庆

# 自我介绍

知乎
www.zhihu.com

知乎技术平台工程师

负责 Kafka 和数据库平台

曾在新浪和金山云负责镜像流量分析项目

# 纲要

知乎
www.zhihu.com

Kafka 在知乎的应用

为什么做基于 Kubernetes 的 Kafka 平台

基于 Kubernetes 的 Kafka 平台实践

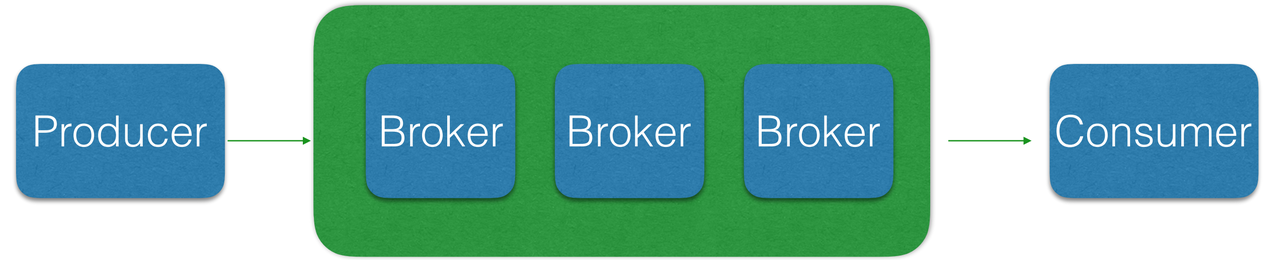
# Apache Kafka

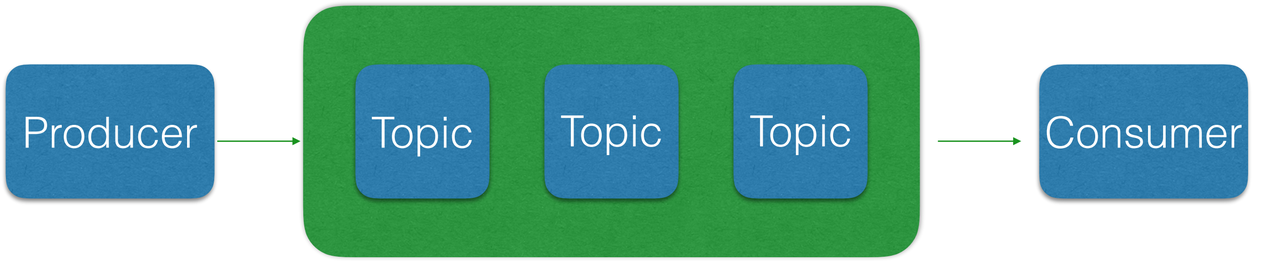
知乎
www.zhihu.com

分布式的流式数据平台

高吞吐

容错性



# Kafka 在知乎的应用

知乎
www.zhihu.com

平台承载知乎业务日志、数据传输和消息队列服务

平台线上稳定运行

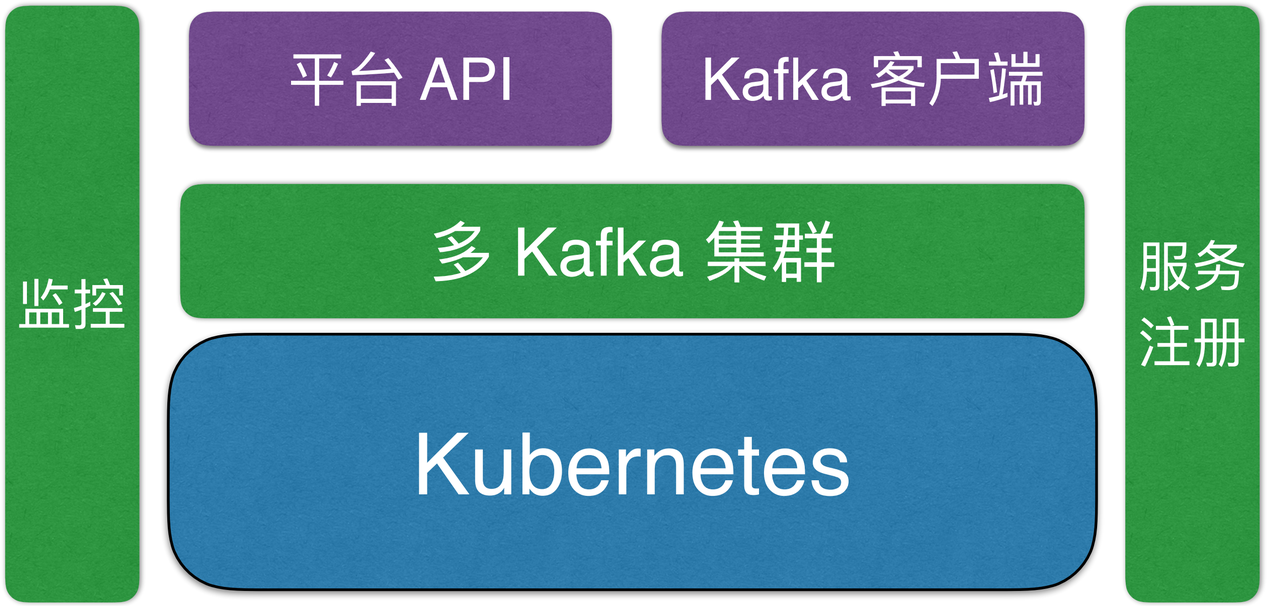
基于 Kubernetes 的 Kafka 集群 13 个， 1000+ Topic

知乎技术平台重要的组件

# 平台概览

知乎
www.zhihu.com

- 多集群
- 高可用



# 为什么采用 Kubernetes

## 问题驱动

- Kafka 资源规划不合理
    - 单一集群造成系统单点
    - 不区分集群和 Topic 等级，影响重要业务
- 业务与 Kafka 深度耦合

# 资源规划

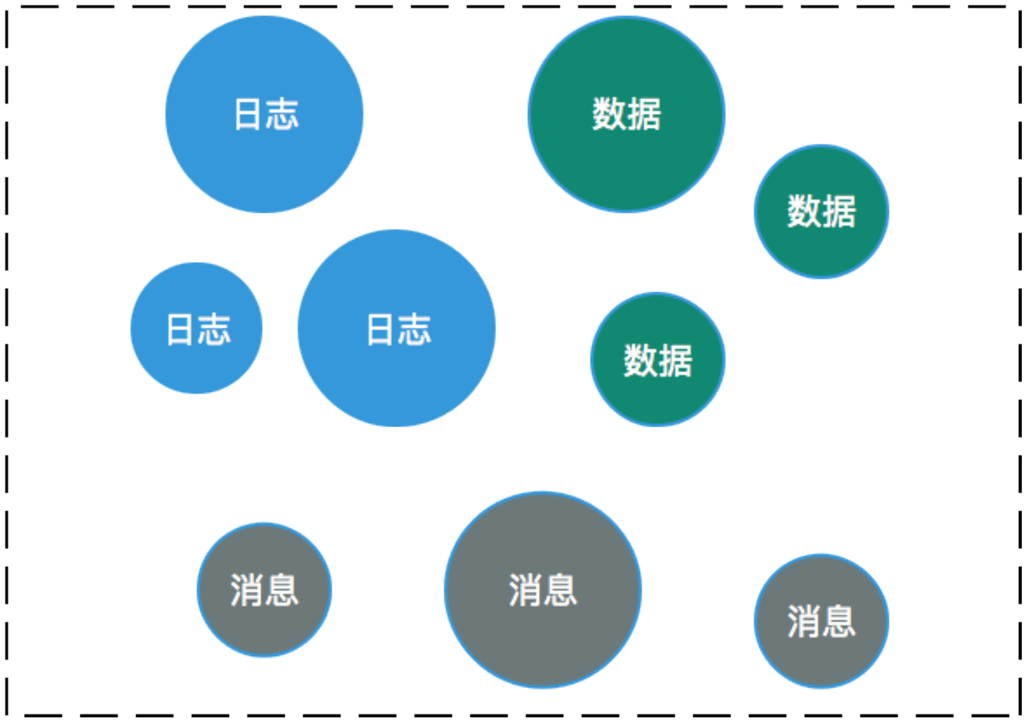
知乎
www.zhihu.com

多 Kafka 集群方式

根据 Topic 类型划分集群

同一类型 Topic 的集群细分

- Topic 服务等级、容量和规模划分



# 多集群问题

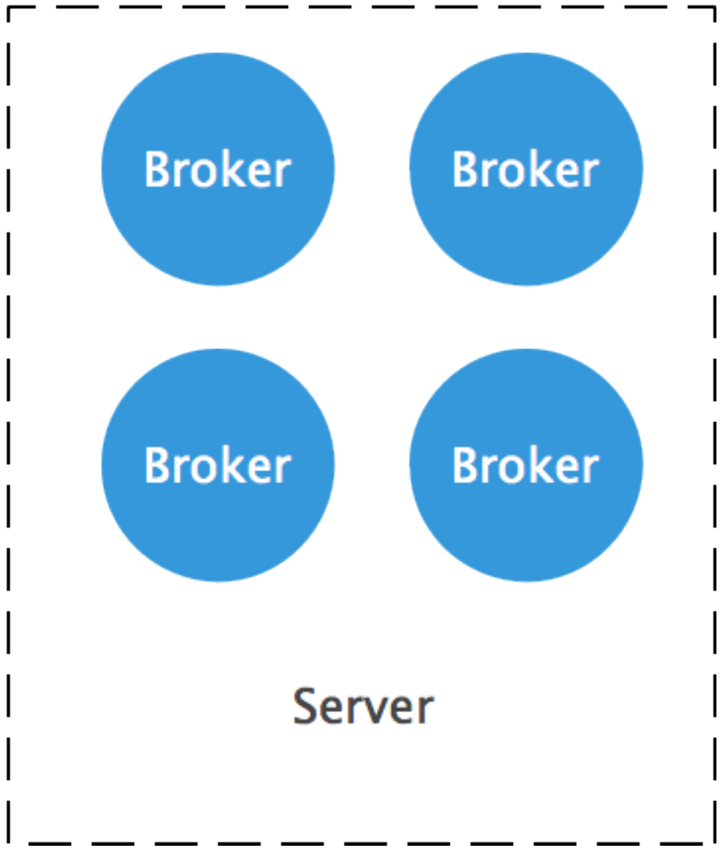


多变需求引发集群规模增长

- Broker, Topic 规模

服务器资源利用率

- 单机运行多 Broker 方式



# 磁盘是个大问题

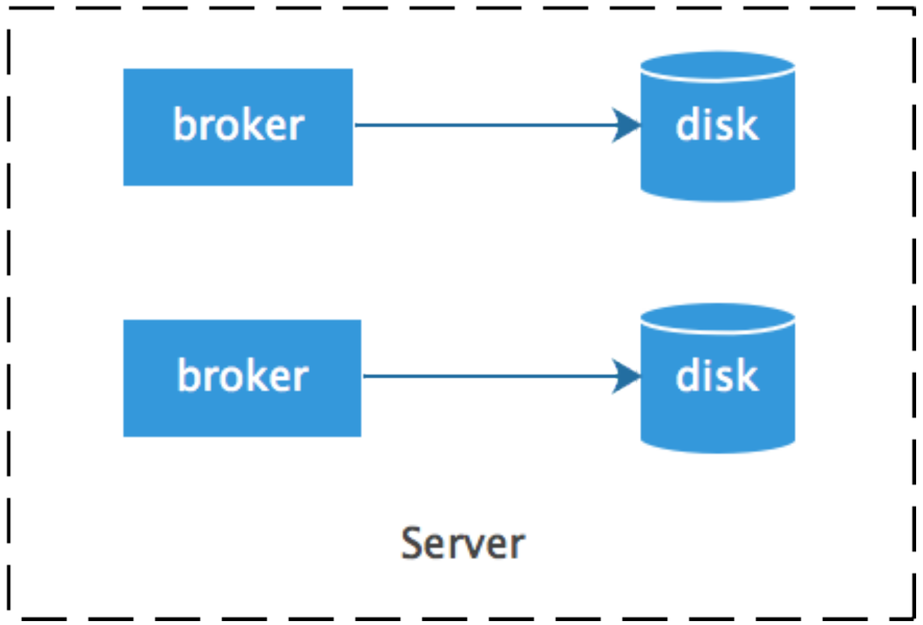
知乎
www.zhihu.com

磁盘是不得不考虑的问题

- 日志落盘，日志失效

方案

- Broker 之间物理磁盘隔离

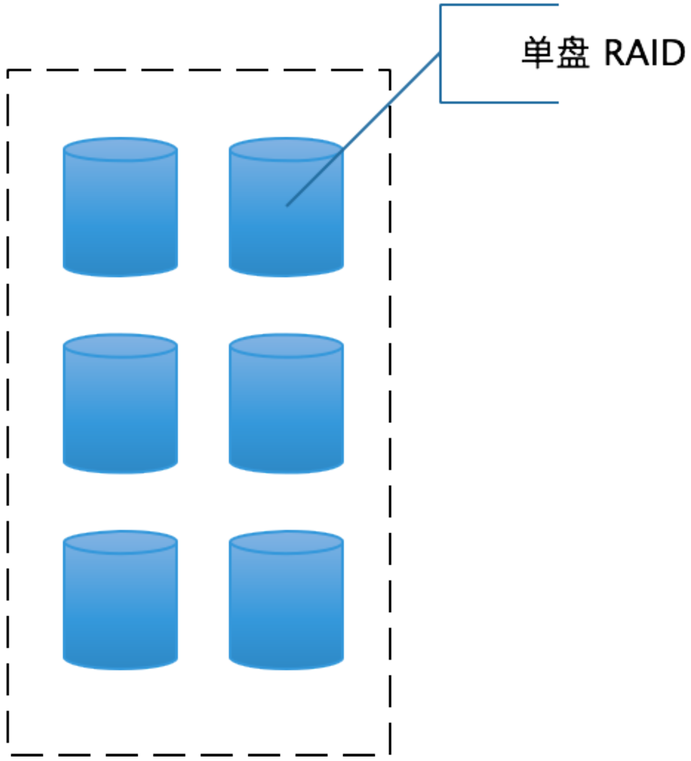


# 服务器选型

知乎
www.zhihu.com

高密度存储服务器

- 多磁盘, 单盘 RAID
- 服务器使用率高



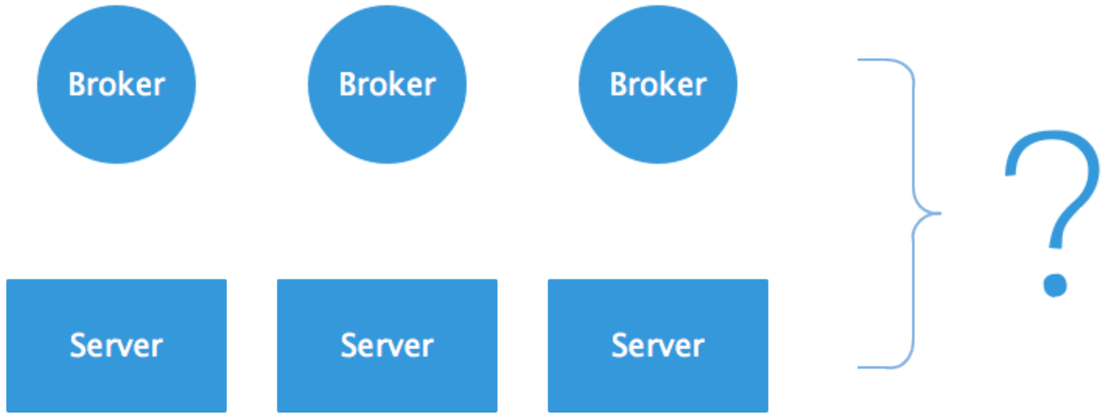
# 如何管理集群

知乎
www.zhihu.com

集群数增加，导致 Broker 扩张

- 如何调度它们
- 如何管理它们

服务器如何管理



# Kubernetes

知乎
www.zhihu.com

集群资源管理和调度

容器技术提供资源隔离

应用程序管理

# Kafka on Kubernetes

知乎
www.zhihu.com

设计 Kafka 容器

- 内存、CPU、网络和存储

调度 Kafka 容器

# 内存 CPU 和网络



内存 和 CPU

- 依照集群类型测试基准数据

容器网络

- 容器采用独立的内网 IP 方案

# 存储

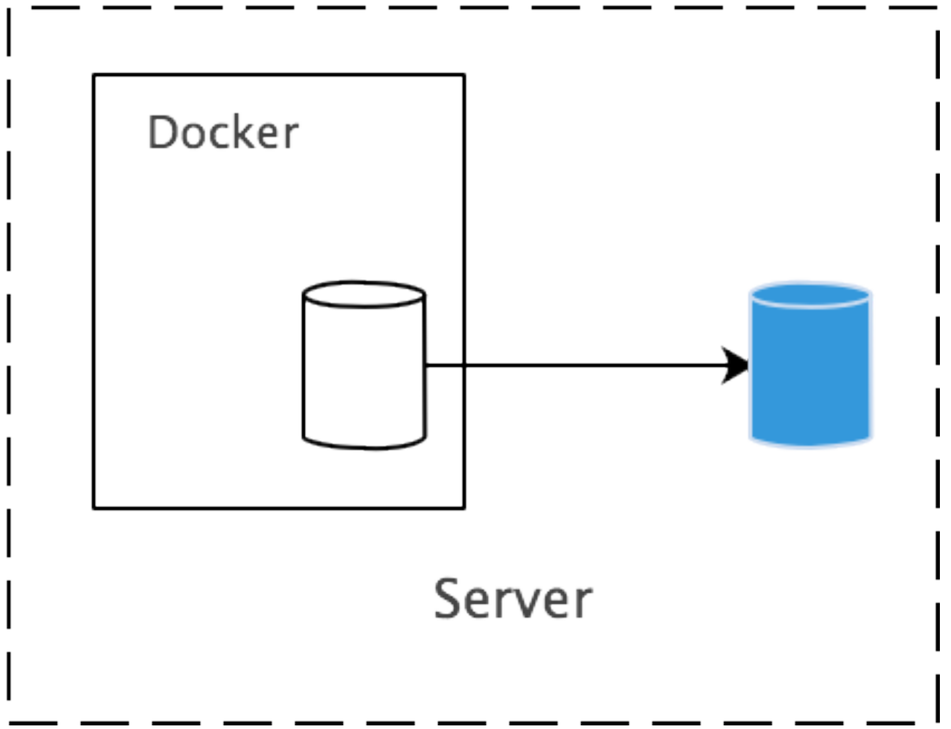


容器挂载服务本地目录

Kafka 高性能

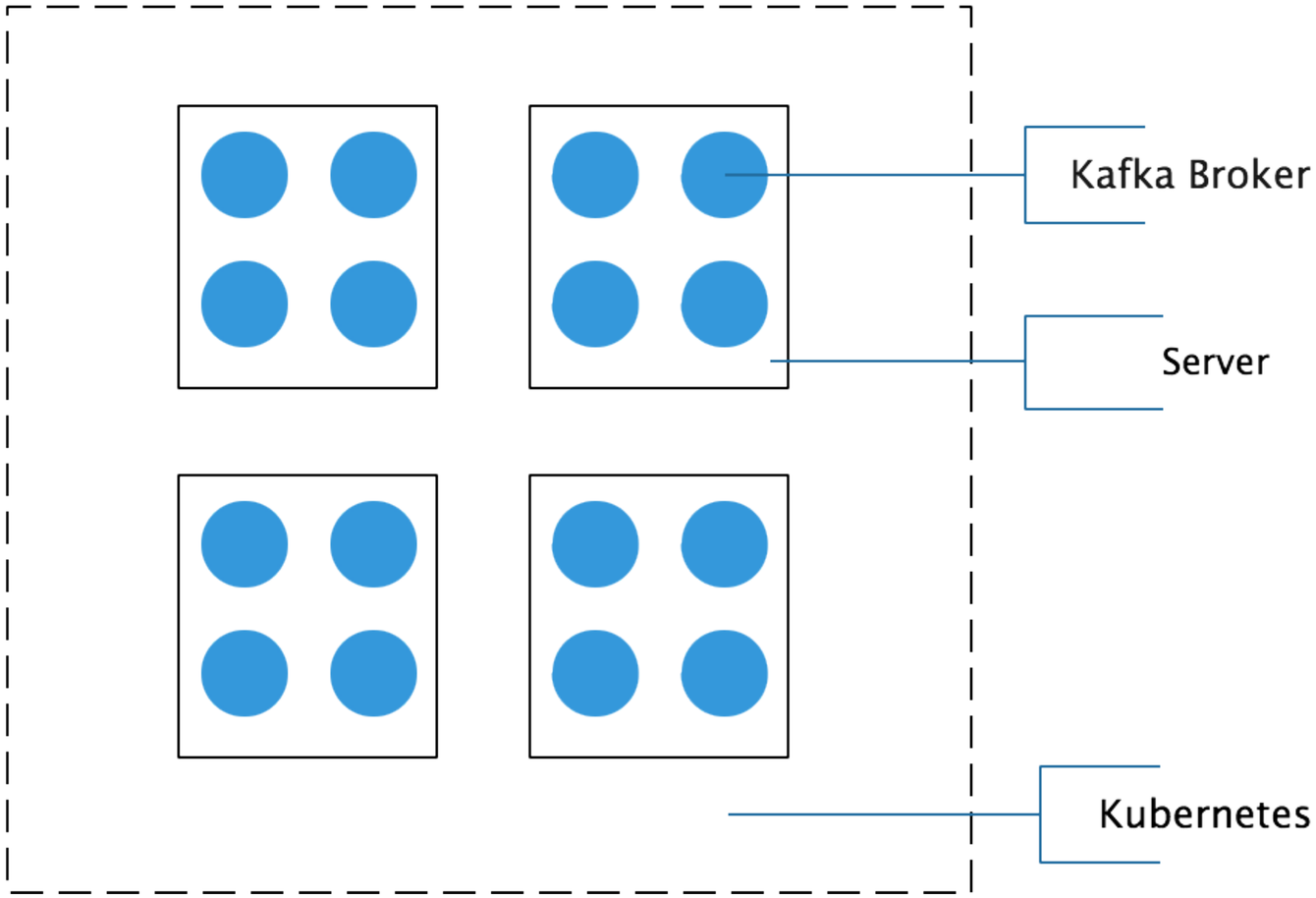
- 文件系统缓存

Kafka 日志落盘



# 集群概览





# 如何调度 Kafka 容器



磁盘是容器的调度单元

目标

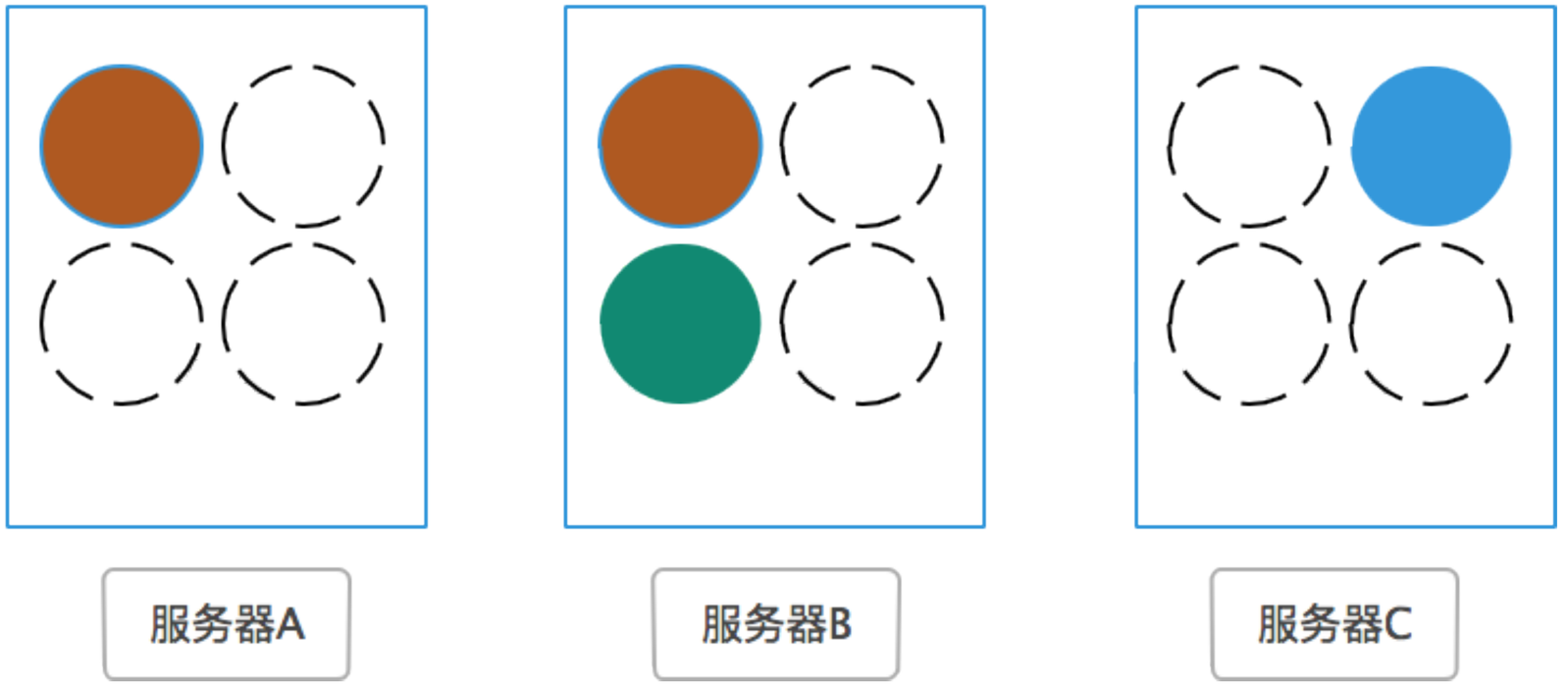
- Broker 在节点分散
- 节点存储使用均匀

# 磁盘调度方法



根据服务器磁盘状态计算分数，分数高者被调度

- 集群 Broker 在节点部署情况
- 服务器可用磁盘情况





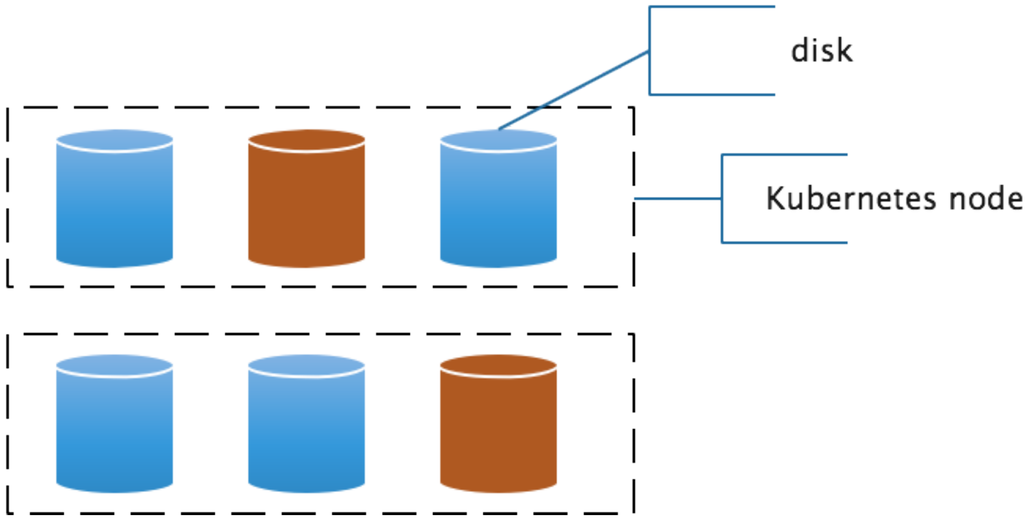

如果创建红色集群则服务器 C 最优

如果创建蓝色集群则服务器 A 最优

# 服务器上线注册磁盘信息

etcd 保存的磁盘信息

- 主机信息，比如 node
- 状态 : unused , used，cleaning
- 其他信息，例如集群信息



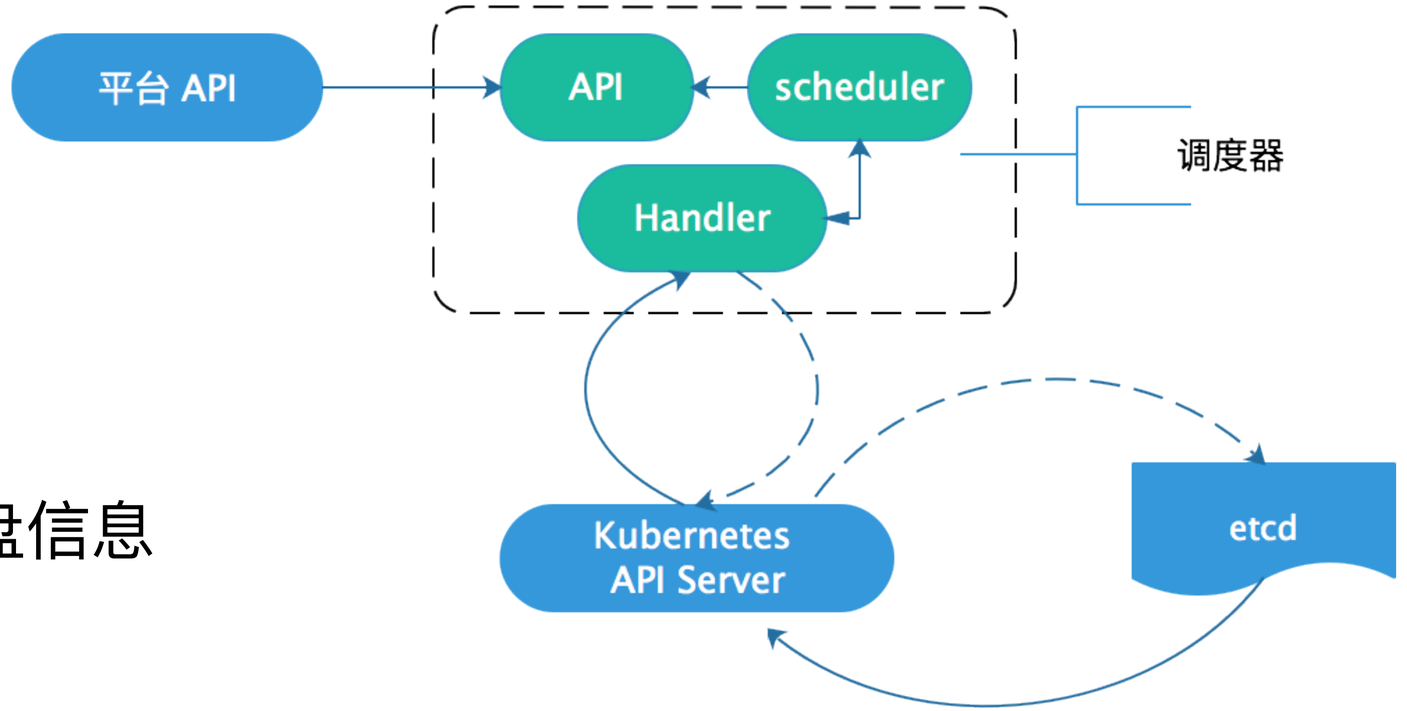
# 磁盘调度器

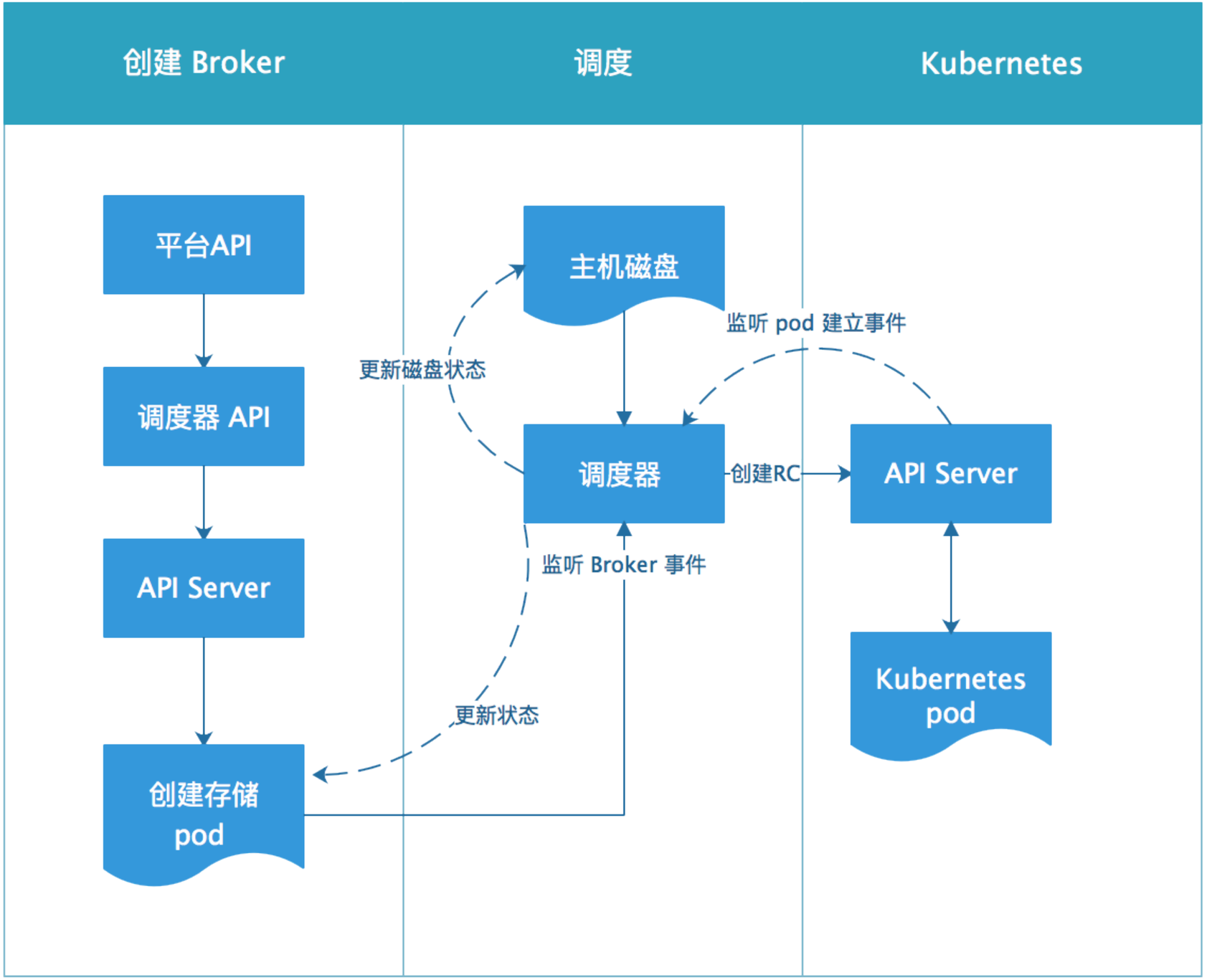
知乎
www.zhihu.com

按照调度算法选择节点

创建 ReplicationController

监听 Kubernetes 状态更新磁盘信息



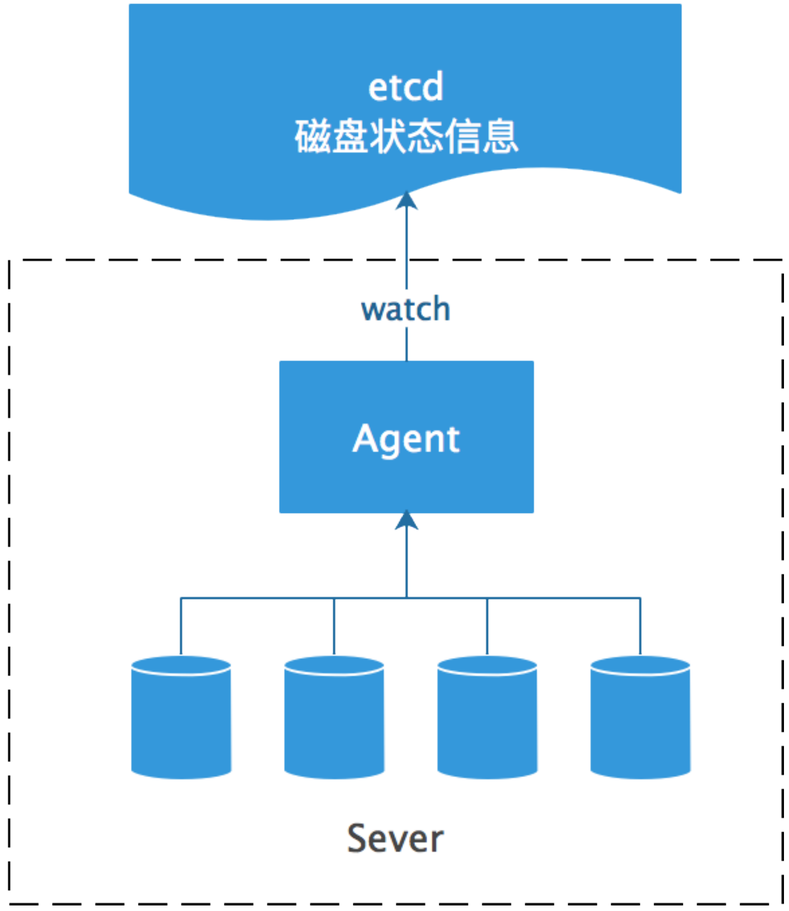


# 本地磁盘管理

知乎
www.zhihu.com

Kubernetes node 部署 Agent

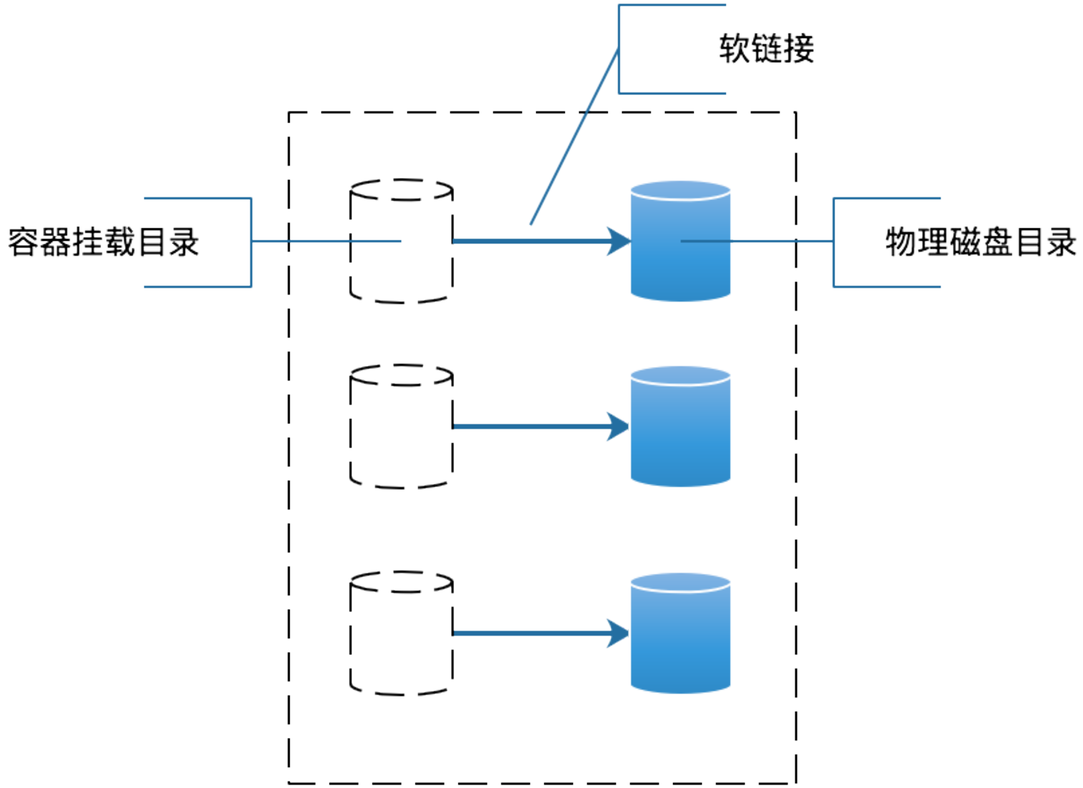
- 监控服务器存储状态
- 磁盘容量资源回收
- 磁盘故障处理



# 本地目录设计



容器挂载磁盘目录的软连接



# 容错

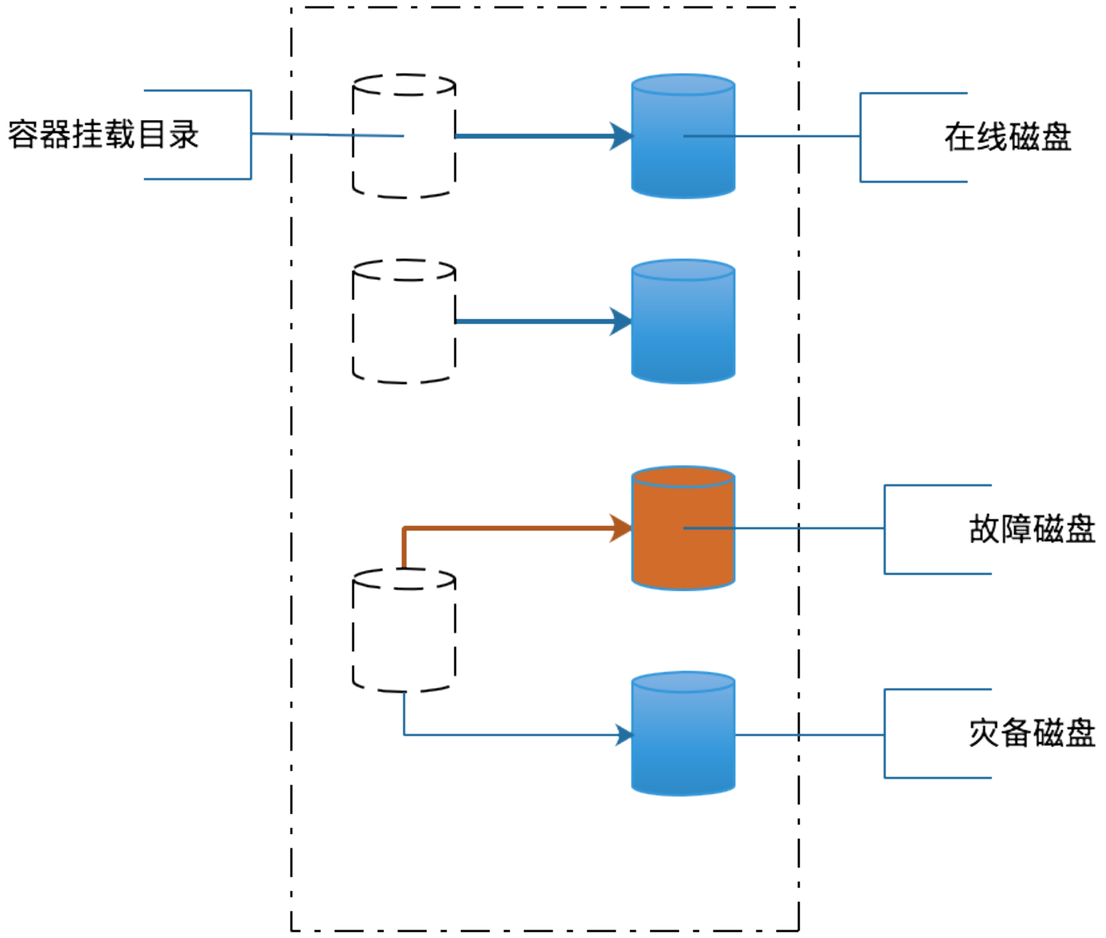


磁盘容错

- 磁盘故障不可避免
- 快速恢复

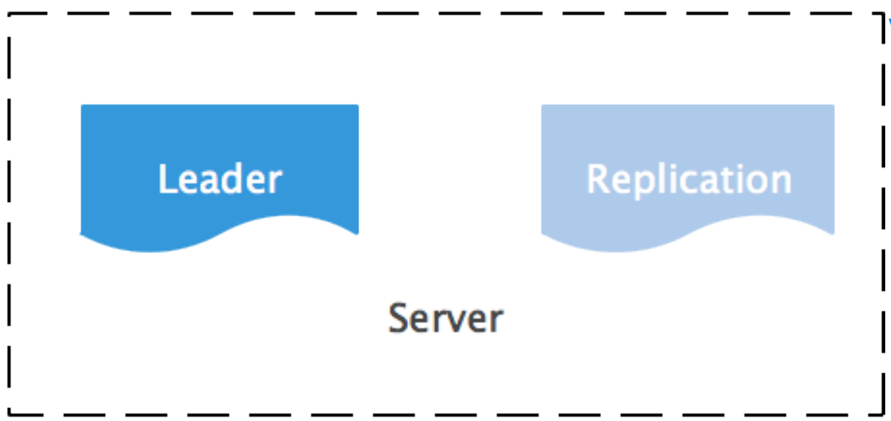
服务器预留备用磁盘

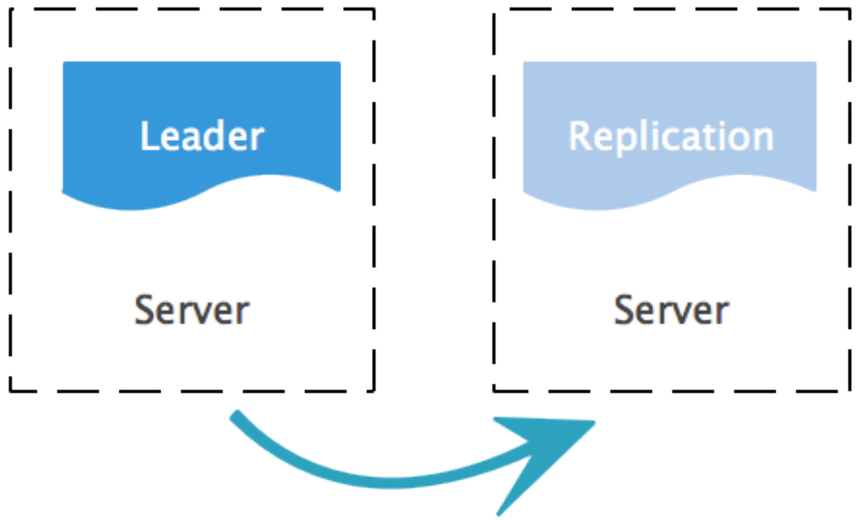
单盘故障启用备用盘



# 主机容错

- 优化磁盘调度算法
- 运用 Kafka 机架感知特性



# 服务注册和客户端

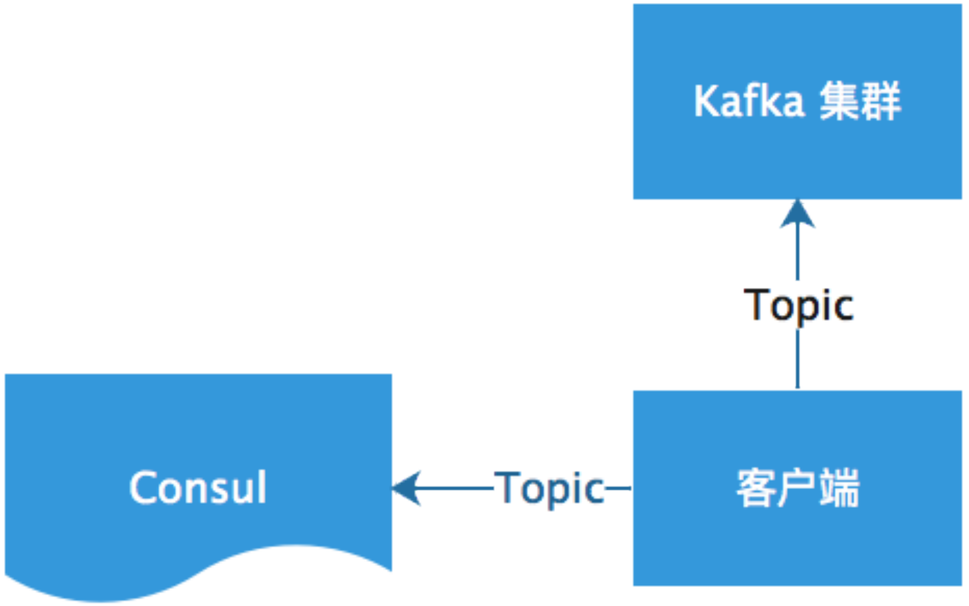


注册 Topic 的集群信息

- Broker, Zookeeper
- Status 是否启用

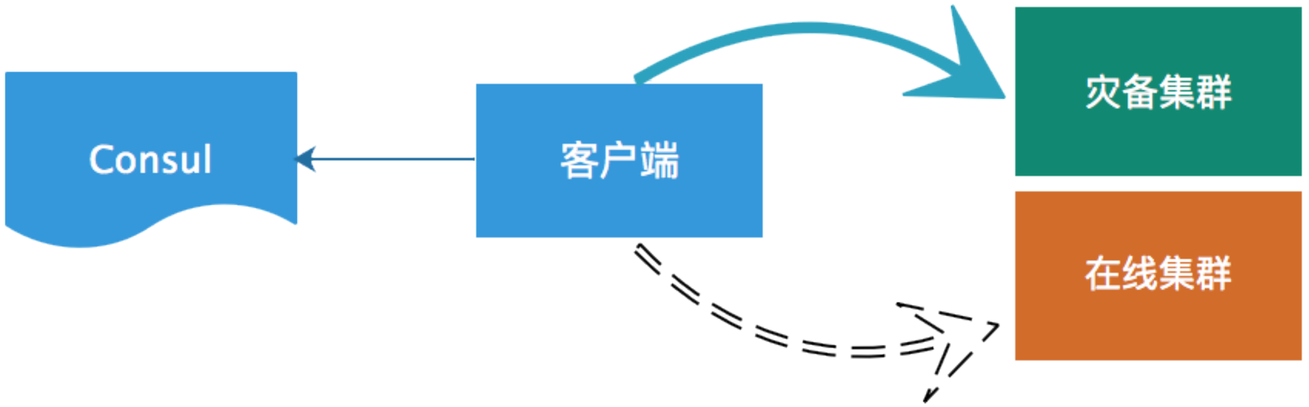
客户端

- 业务易用
- 标准客户端，降低集群风险



# 集群容错

- 灾备集群
  - 保证重要 Topic 高可用
- 客户端与服务器注册联动



# 监控



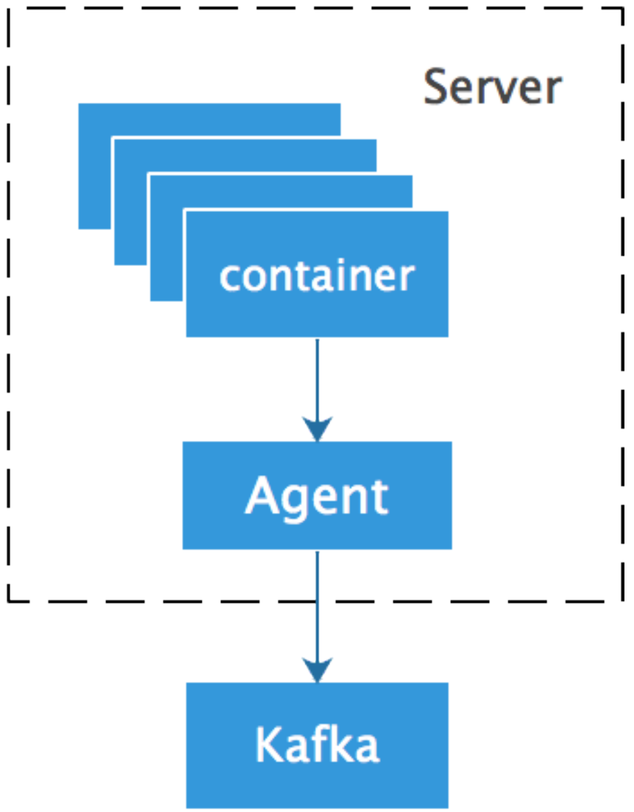
| | 指标维度 | 举例 |
|---|---|---|
| **Kubernetes** | 3 | 容器内存、CPU、运行状态 |
| **Broker** | 14 | 消息量，JVM,<br>Leader分布，磁盘消耗 |
| **Topic** | 13 | 消息量，消费延迟 |
| **主机** | 4 | 内存、网络、CPU、磁盘 |
| **客户端** | 2 | 生产或消费 Topic 消息量 |

# 业务解耦

知乎
www.zhihu.com

容器日志通过本地代理收集

- 输出重定向



# 未来

知乎
www.zhihu.com

# 谢谢