OpenEuler

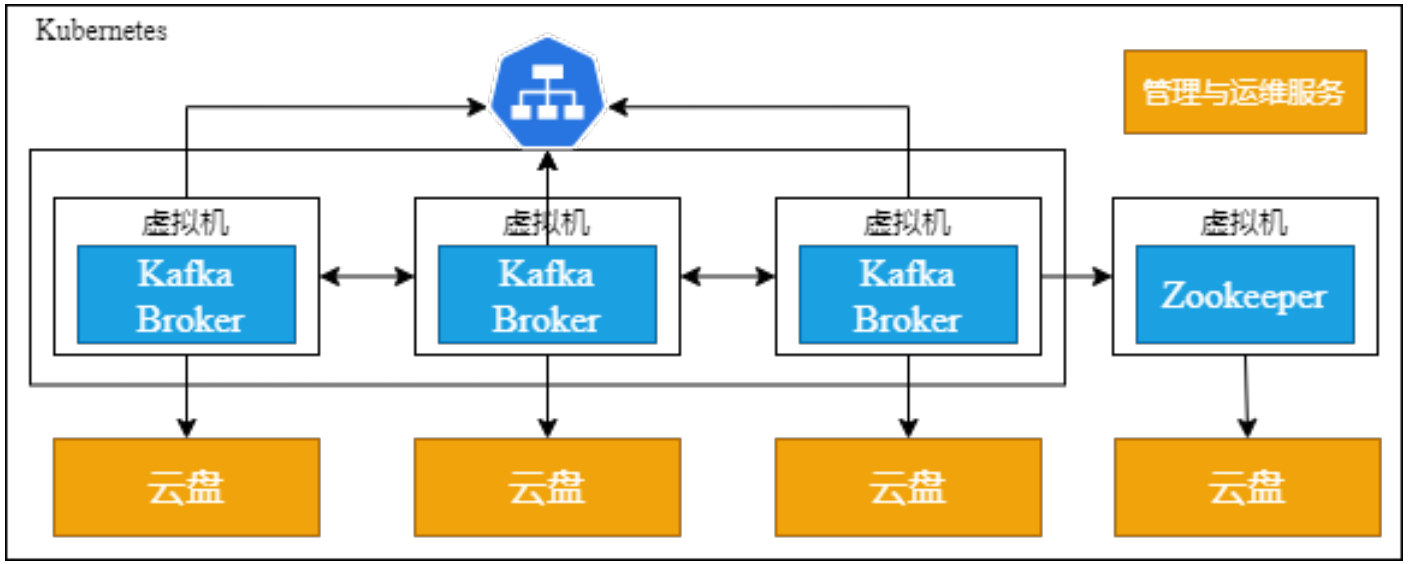
# 移动云Kafka的云原生架构演进之路

中国移动云能力中心

王嘉凌

# 目录

# 基于虚拟机架构的Kafka架构



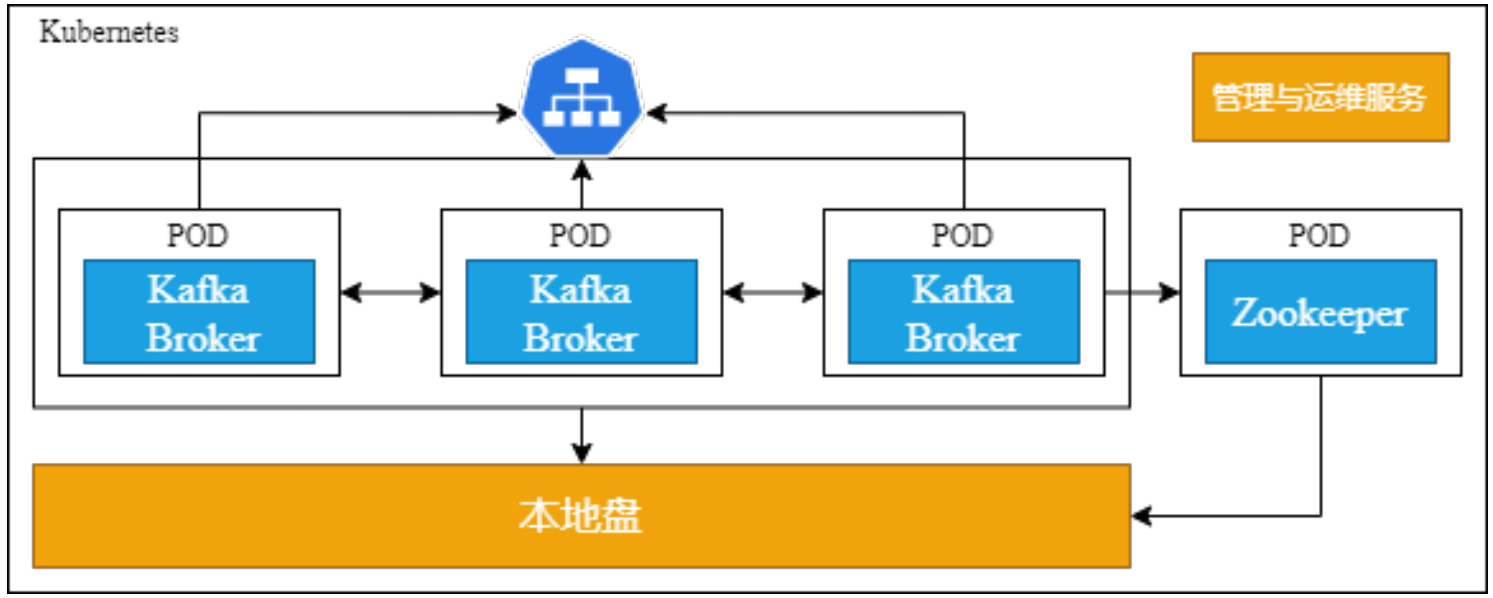

基于虚拟机架构的移动云Kafka痛点

1）虚拟机迁移成本高
2）存储采用云盘，性能和本地盘相比有较大差异
3）算存一体架构导致的扩容困难，运维成本高等

云原生Kafka架构升级的迫切需求

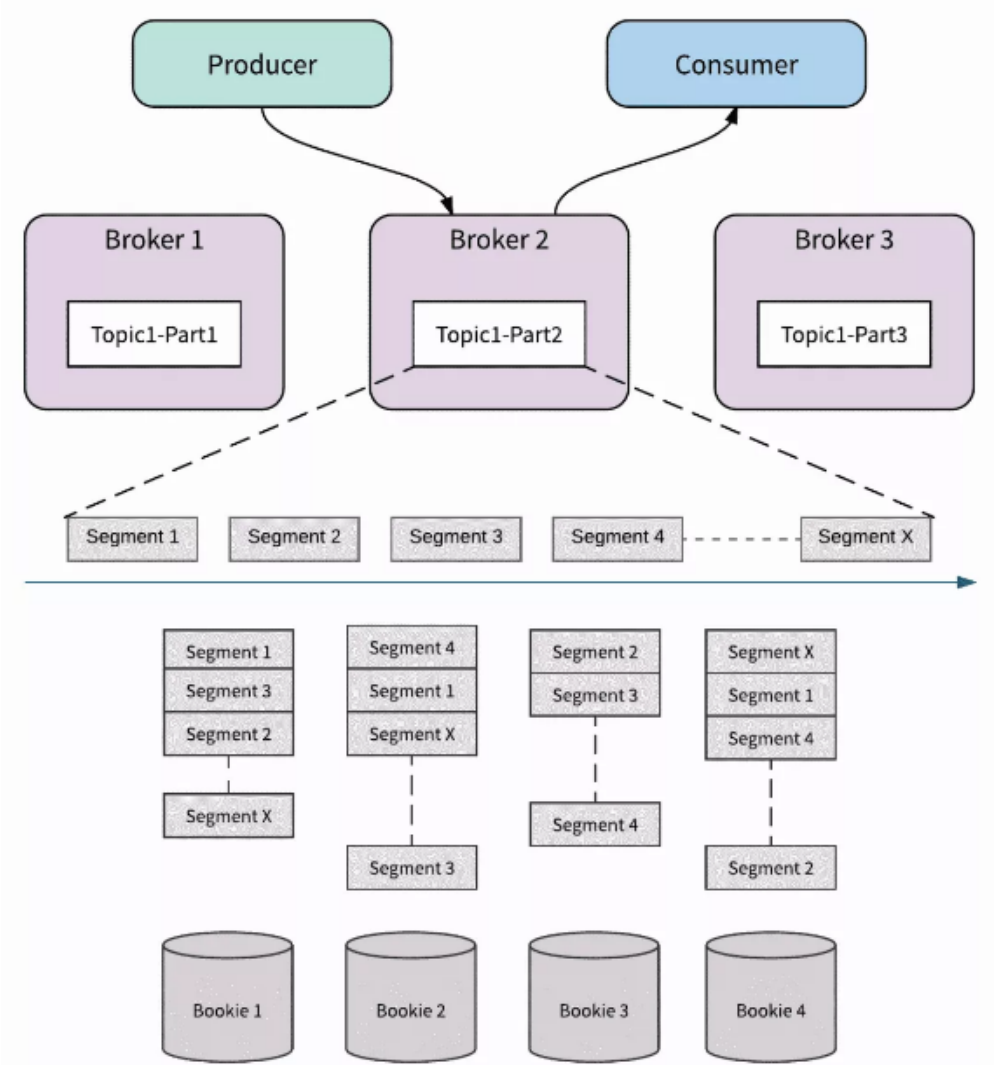
1）采用本地盘，高速缓存提高性能
2）高效编排，动态扩缩
3）支持跨可用区灾备
4）集约化架构 - 云化集中式管理

# 基于K8S的Kafka架构




1）多个Kafka集群共享使用一个本地存储盘，无法发挥Kafka顺序写盘的性能优势
2）存算一体架构扩展性较低（数据分区多副本管理）
3）租户存储网络隔离功能缺失

# Pulsar特性 - 计算存储分离




1）计算存储分离

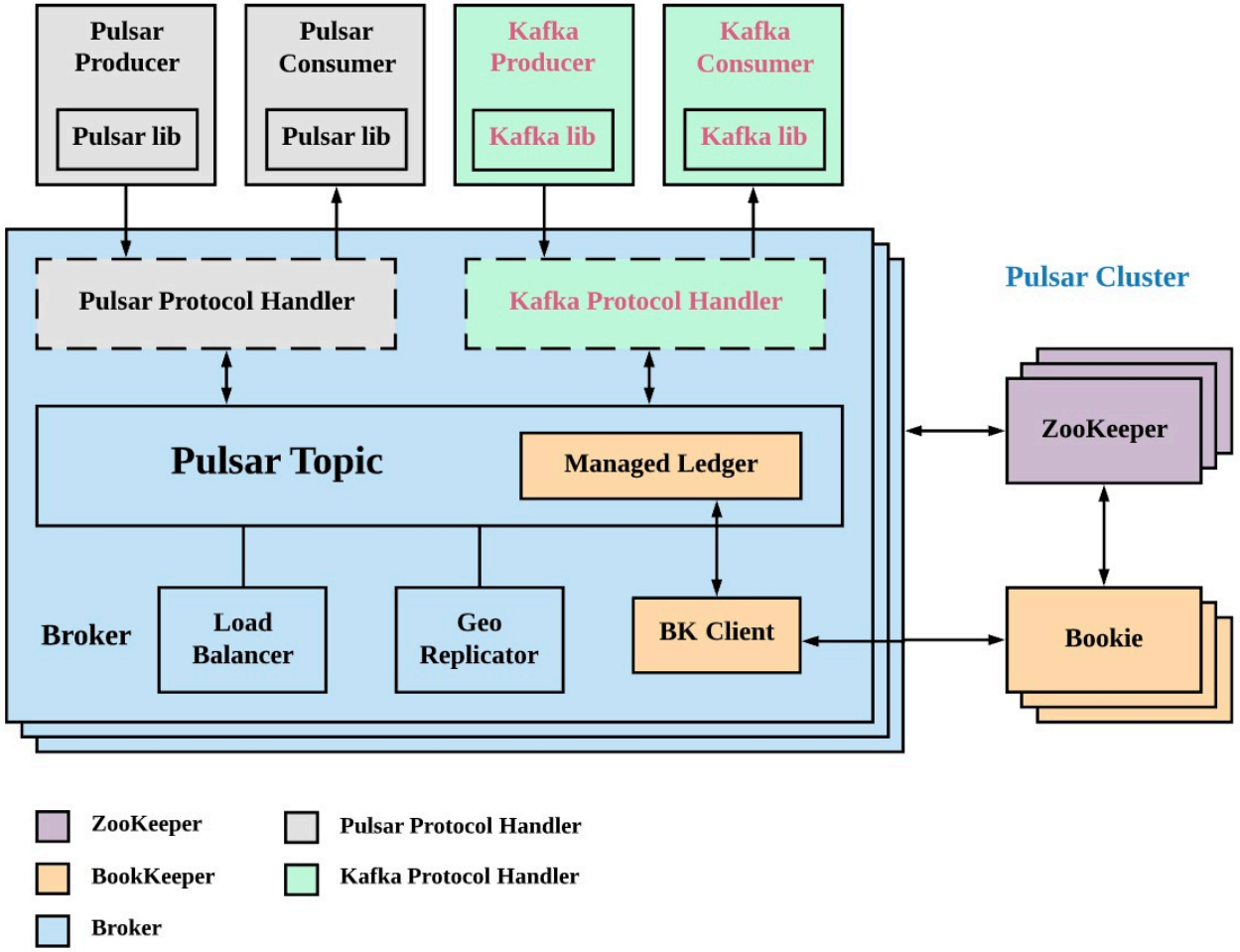
2）底层存储使用BookKeeper，单个Topic的数据可以分散存储在多个Bookie节点上，支持磁盘顺序读写

3）实时水平扩容计算和存储能力以及Topic分区

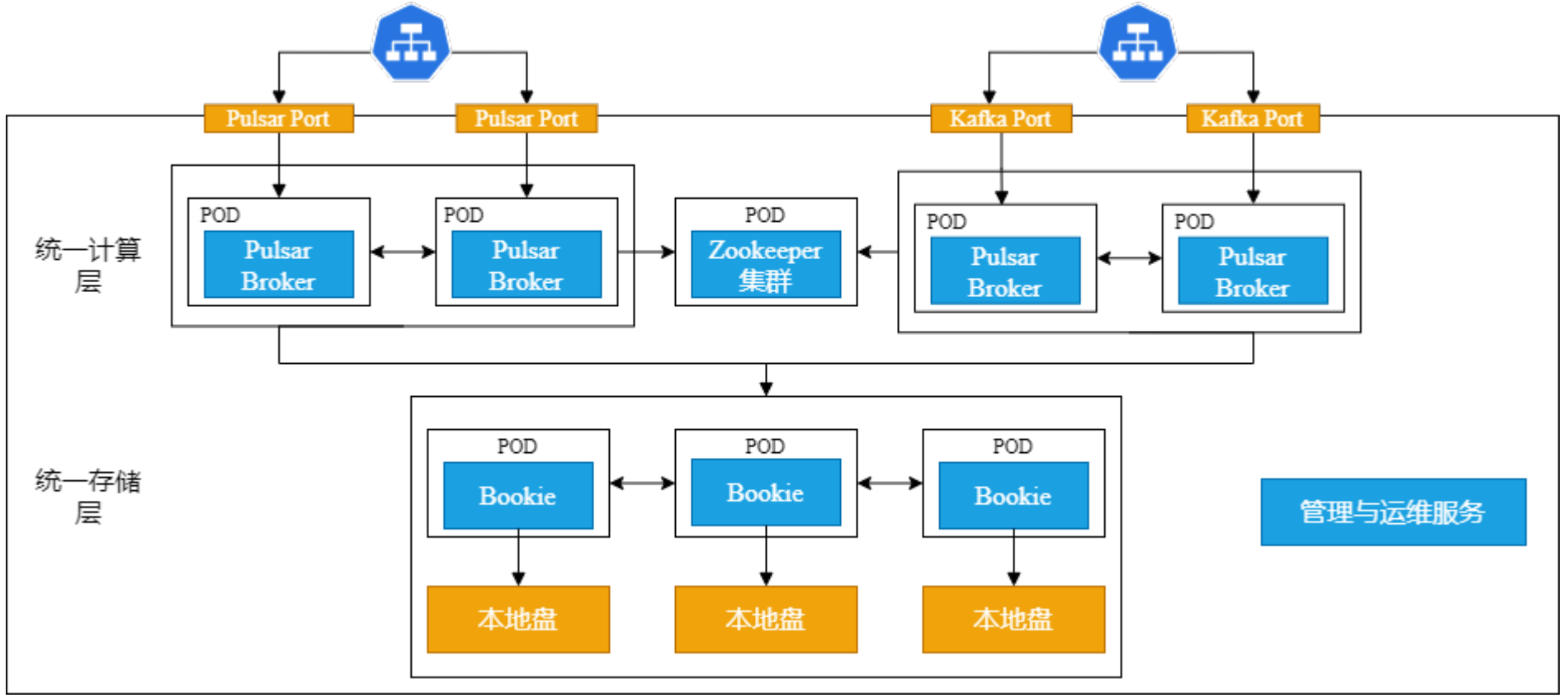
4）高可用的集群容灾

# Kafka On Pulsar

基于Apache Pulsar提供的Protocol Handler接口实现的KoP，可以让Pulsar具有处理Kafka客户端请求的能力



# 基于Pulsar和KoP的Kafka云原生方案




1）架构优势：计算存储分离

2）IO性能优势：底层使用共享BookKeeper集群，可发挥顺序写盘的优势

3）运维优势：复用移动云Pulsar的底层资源布局和管理运维服务

# 方案比较

OpenEuler

| 功能项 | 开源Kafka容器化 | 基于KoP的云原生Kafka |
| --- | --- | --- |
| 性能 | 无法发挥磁盘顺序读写的特性 | KoP内部需要做协议转换，消耗额外的计算资源<br>计算存储分离架构消耗额外的内网带宽 |
| 功能 | 和开源Kafka一致 | KoP存在与原生Kafka不兼容或支持不完善的功能 |
| 架构 | 计算存储不分离带来一系列的痛点 | 计算存储分离，且具有租户隔离，跨区域高可用等特性 |
| 成熟度 | 成熟度高，用户多 | 用户相对不多，主要用于从Kakfa迁移到Pulsar时的过渡方案 |
| 成本 | 运维成本高 | 复用移动云Pulsar的研发和运维能力 |

# 云原生场景下的租户隔离

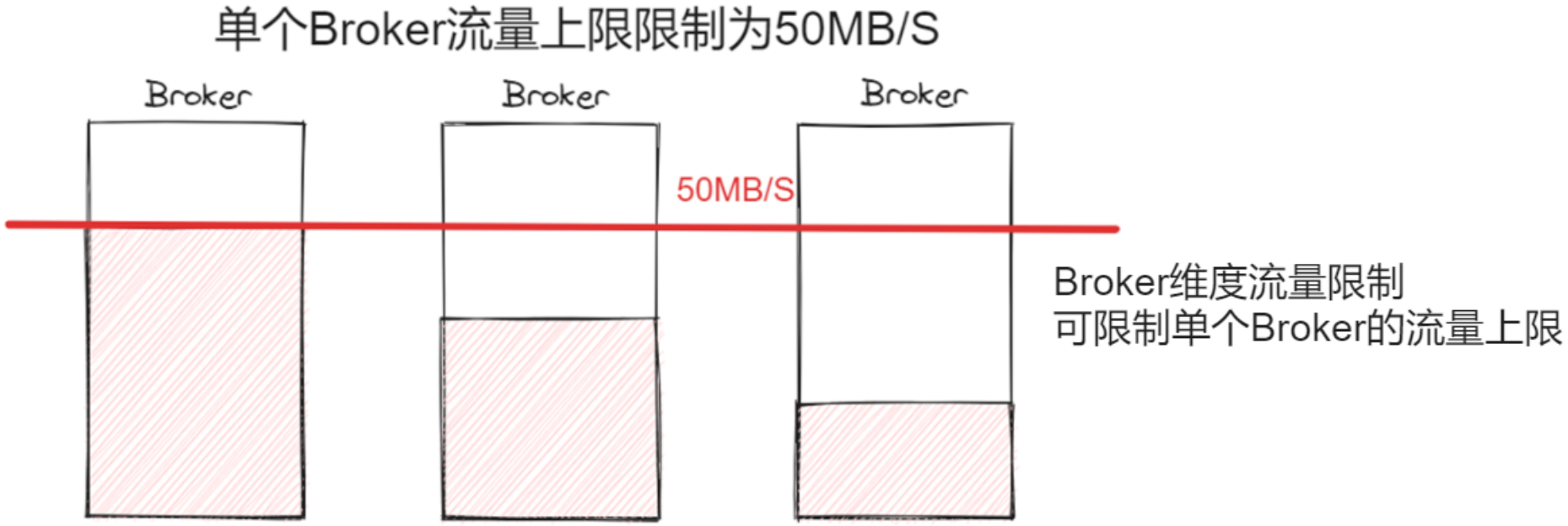
OpenEuler

- **多租户，云原生场景下对消息队列的隔离需求**
  - 计算（CPU，内存），网络（流量上限），存储（磁盘使用量）维度的资源隔离
  - 在单个集群具备控制流量上限和存储使用上限的能力，并且可动态调整

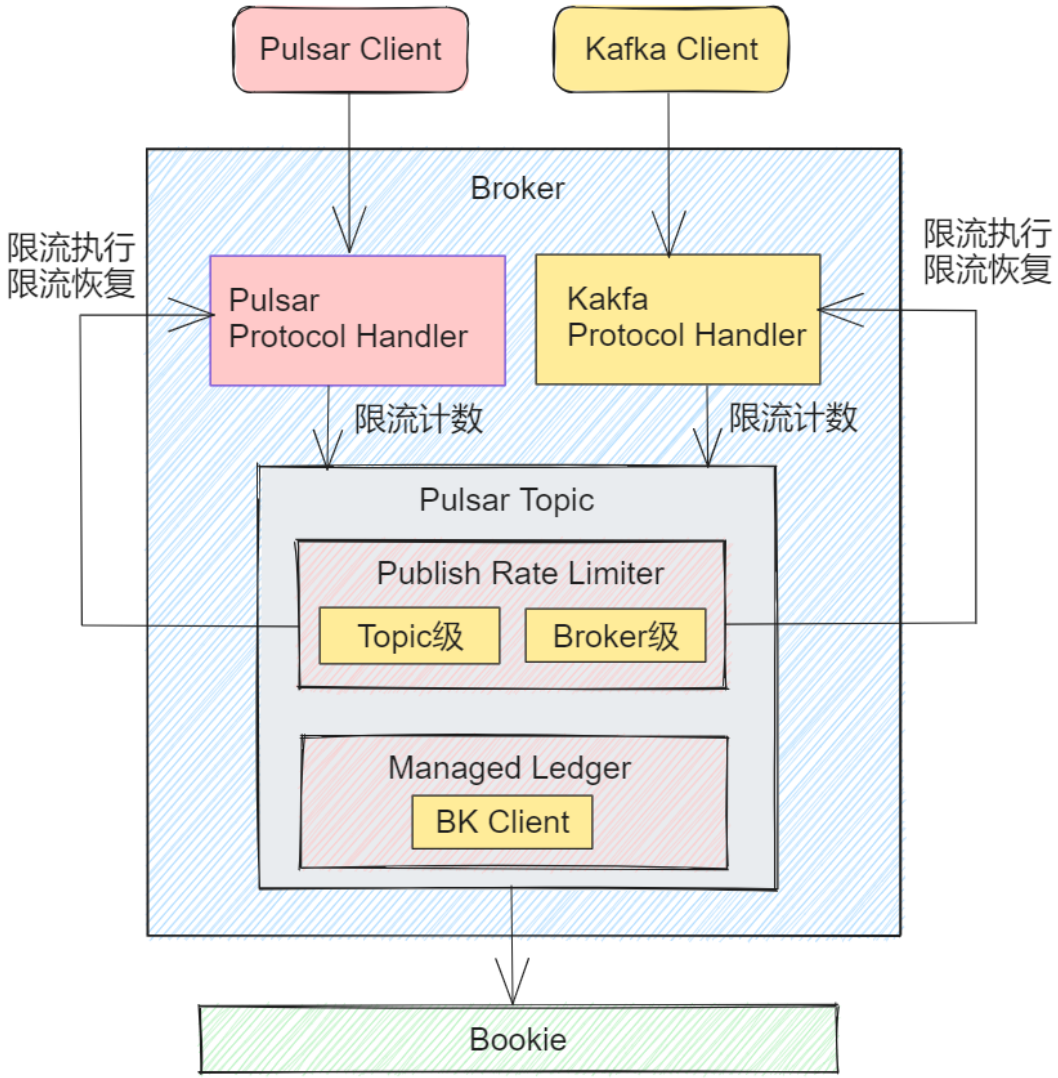
- **多租户隔离实现方案**
  - 计算（CPU，内存）： K8s天然支持
  - 存储 ：通过CSI存储插件实现
  - 网络 ：通过Broker维度的流量限制实现

# Broker维度流量限制的缺陷

- **无法准确限制集群整体的流量上限**
  - Broker之间负载分布不均衡的场景下，只有部分Broker达到限流上限
  - 集群总体流量未能达到预期



# Pulsar流量限制实现原理




- **限流计数：**

往Topic中写入生产数据时，向各个限流器的计数器累加写入的数据量，即消息数和消息大小

- **限流执行：**

限流计数器达到上限时触发Protocol Handler执行限流，停止从Producer接收数据

- **限流恢复：**

到达下一个限流器周期后触发Protocol Handler限流恢复，计数器清零，开始从Producer接收数据

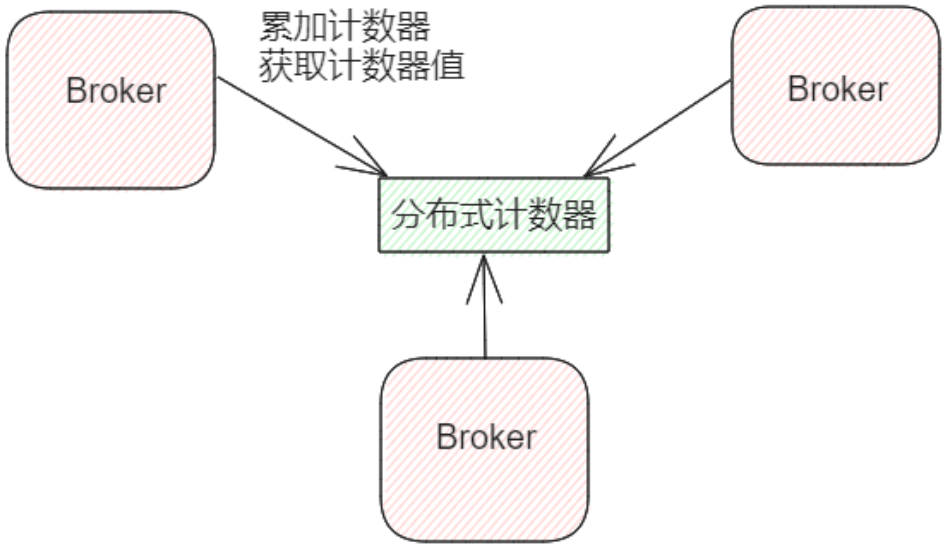
# 集群维度的限流思路

OpenEuler

- **难点**

Broker维度的限流，限流器所需限制的topic都在同一个broker中，计数器可以使用内存变量。

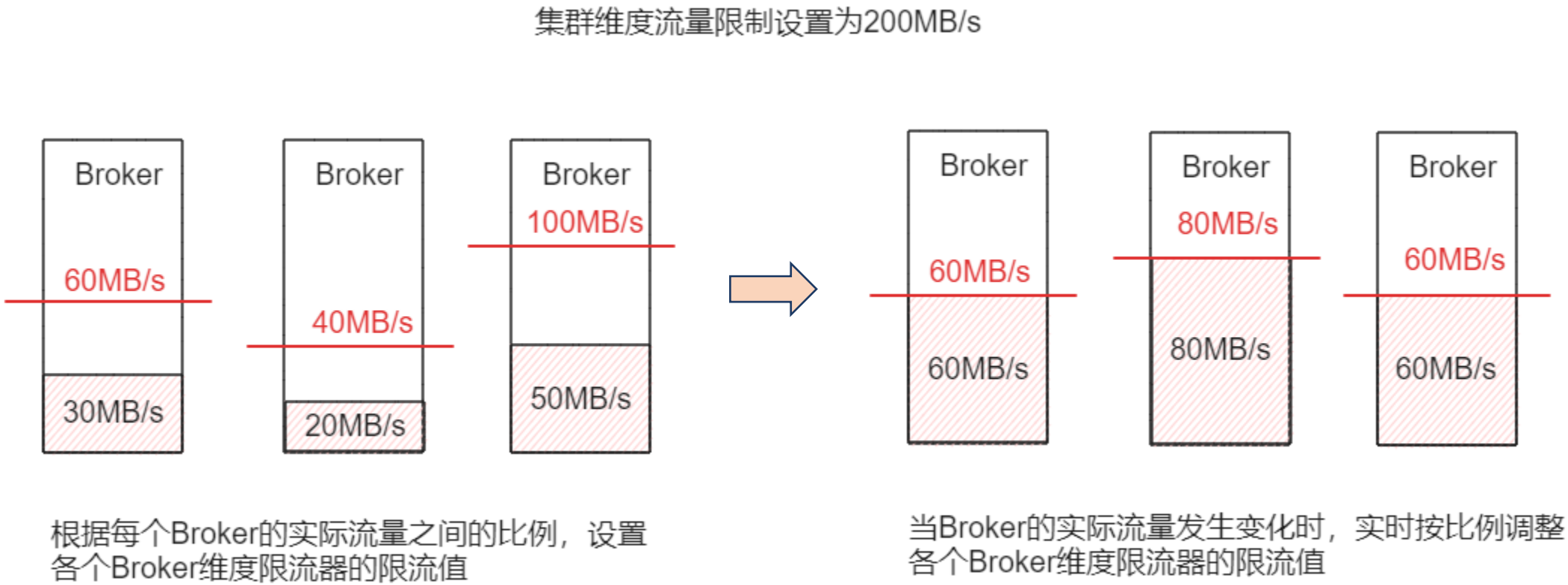
而集群维度的限流，需要限制的topic分布在不同broker中，需要用全局的限流器（分布式缓存）来实现。



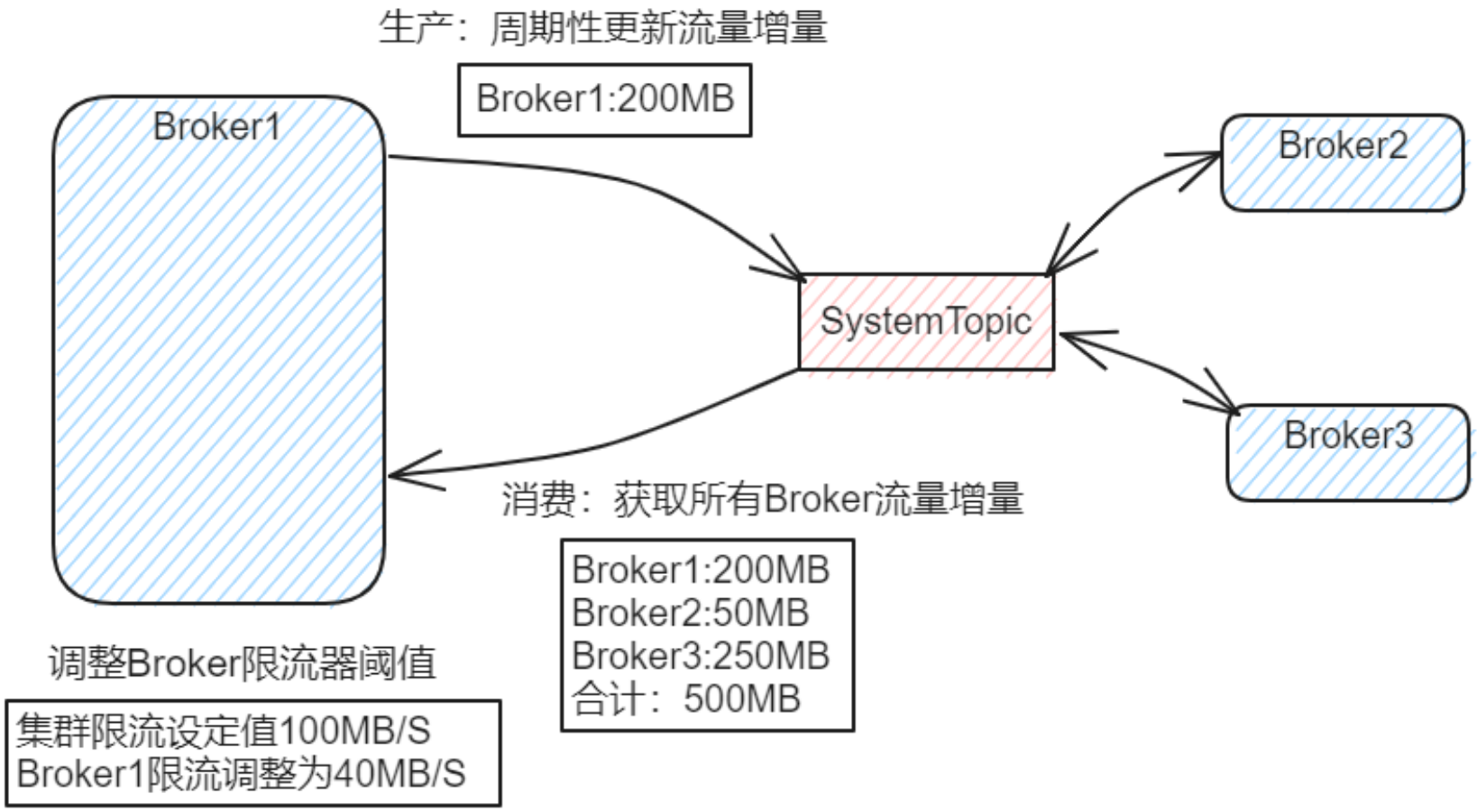

- 影响集群TPS性能
- 需要额外的成本去维护一套分布式缓存

# 集群维度的限流方案

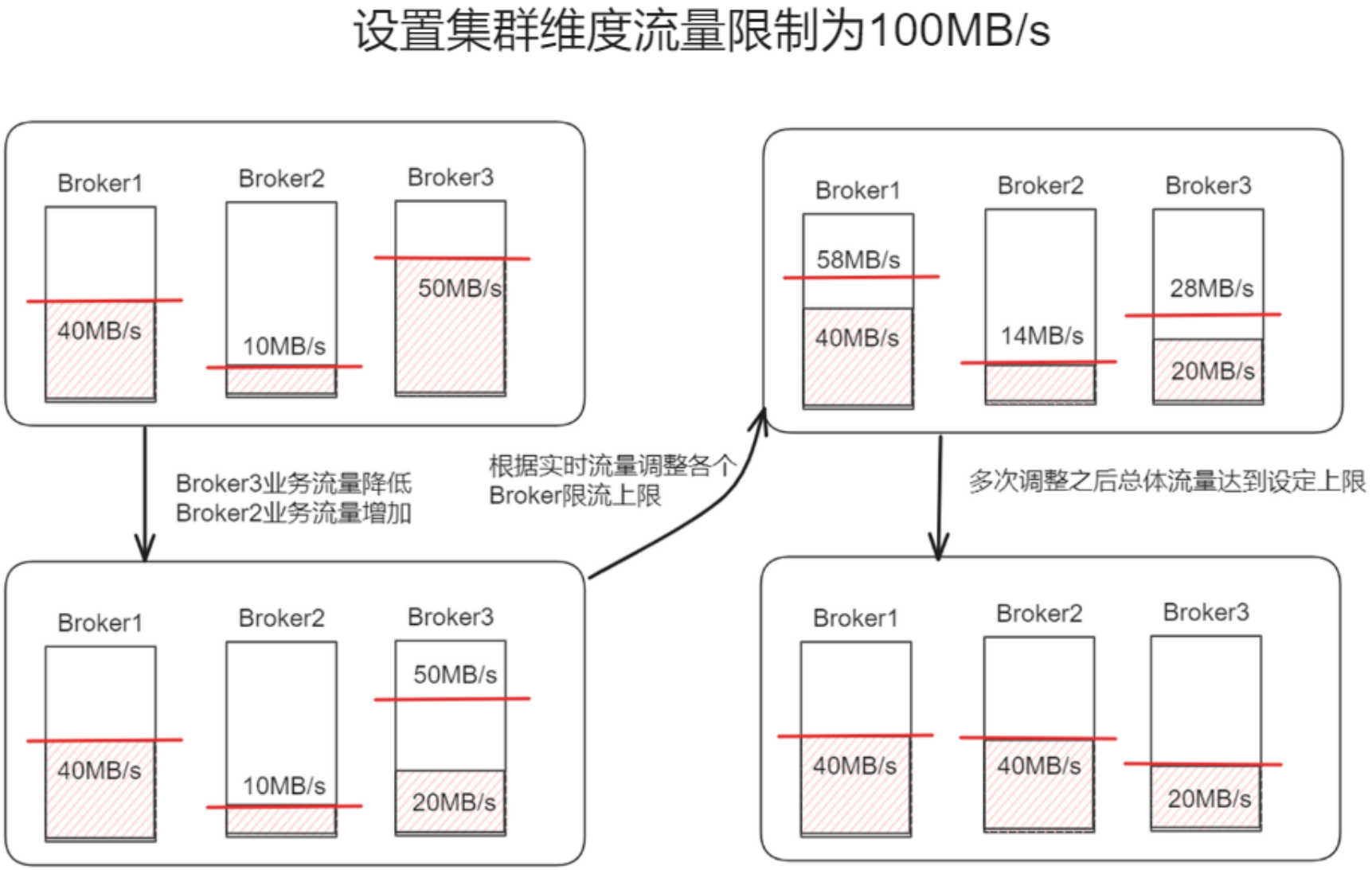
- 基于Broker维度的限流，并根据业务流量的分布，动态调整各个Broker的限流值

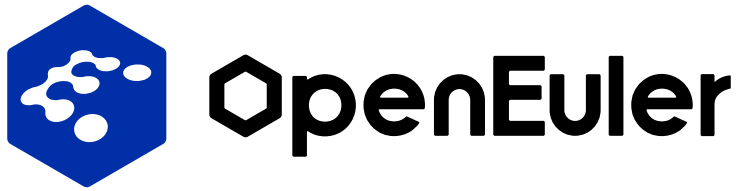


# 集群维度的限流实现


生产：周期性更新流量增量
Broker1:200MB
Broker1
Broker2
SystemTopic
Broker3
消费：获取所有Broker流量增量
Broker1:200MB
Broker2:50MB
Broker3:250MB
合计：500MB
调整Broker限流器阈值
集群限流设定值100MB/S
Broker1限流调整为40MB/S

# 集群维度的限流动态调整效果示意




OpenEuler