Apache Pulsar在腾讯云上的大规模实践

韩明泽@腾讯

王震江@腾讯

王震江

- 腾讯研发工程师，负责腾讯云TDMQ For Pulsar商业化开发
- 开源社区爱好者

韩明泽

- 腾讯高级工程师，负责腾讯云TDMQ For Pulsar商业化开发
- 拥有7年消息队列开发经验, 熟练掌握Pulsar、Kafka、RocketMQ等主流消息队列
- Apache Pulsar/BookKeeper/Zookeeper contributor, RoP maintainer

1.多网接入

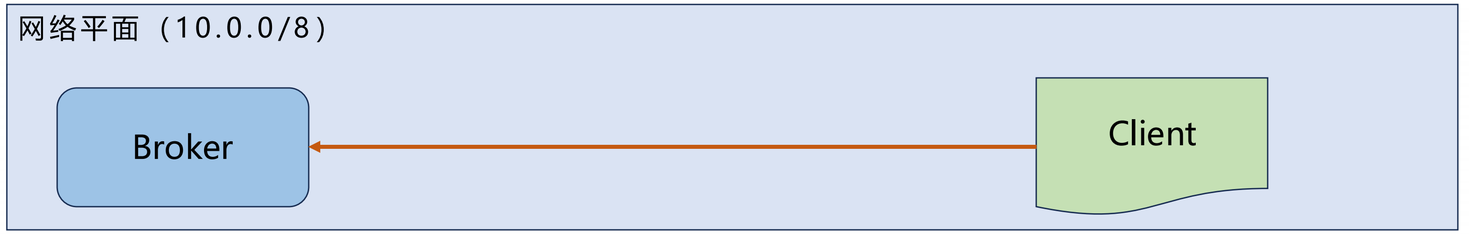
2.集群迁移

3.高可用最佳实践

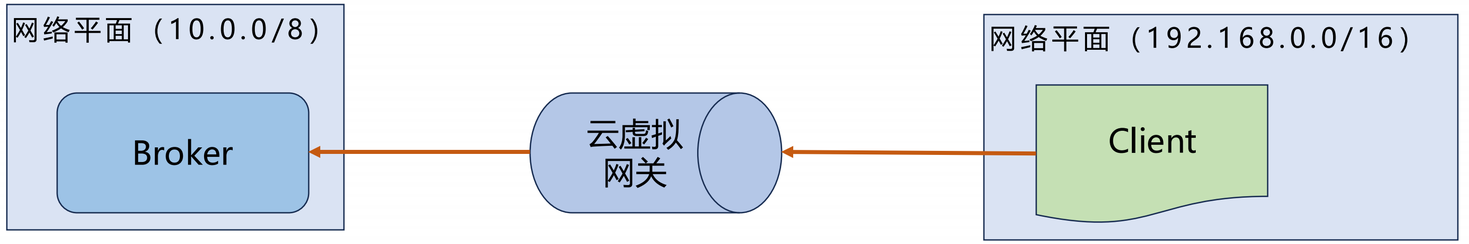
1. 多网接入-网络介绍

- 网络打通
- IP地址映射

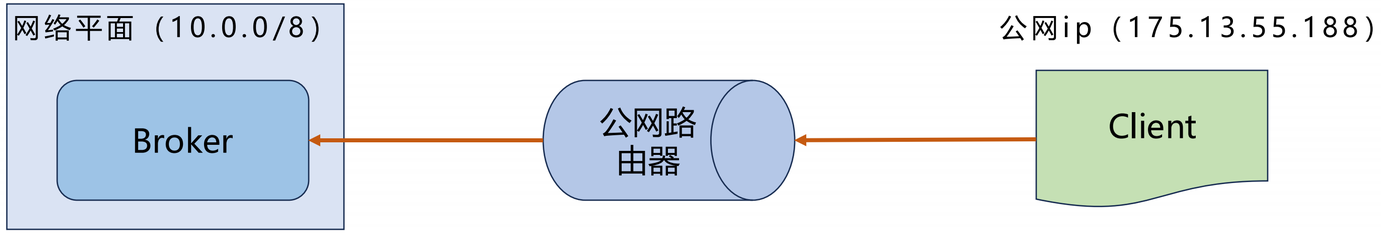
内网



VPC



公网



1. 多网接入

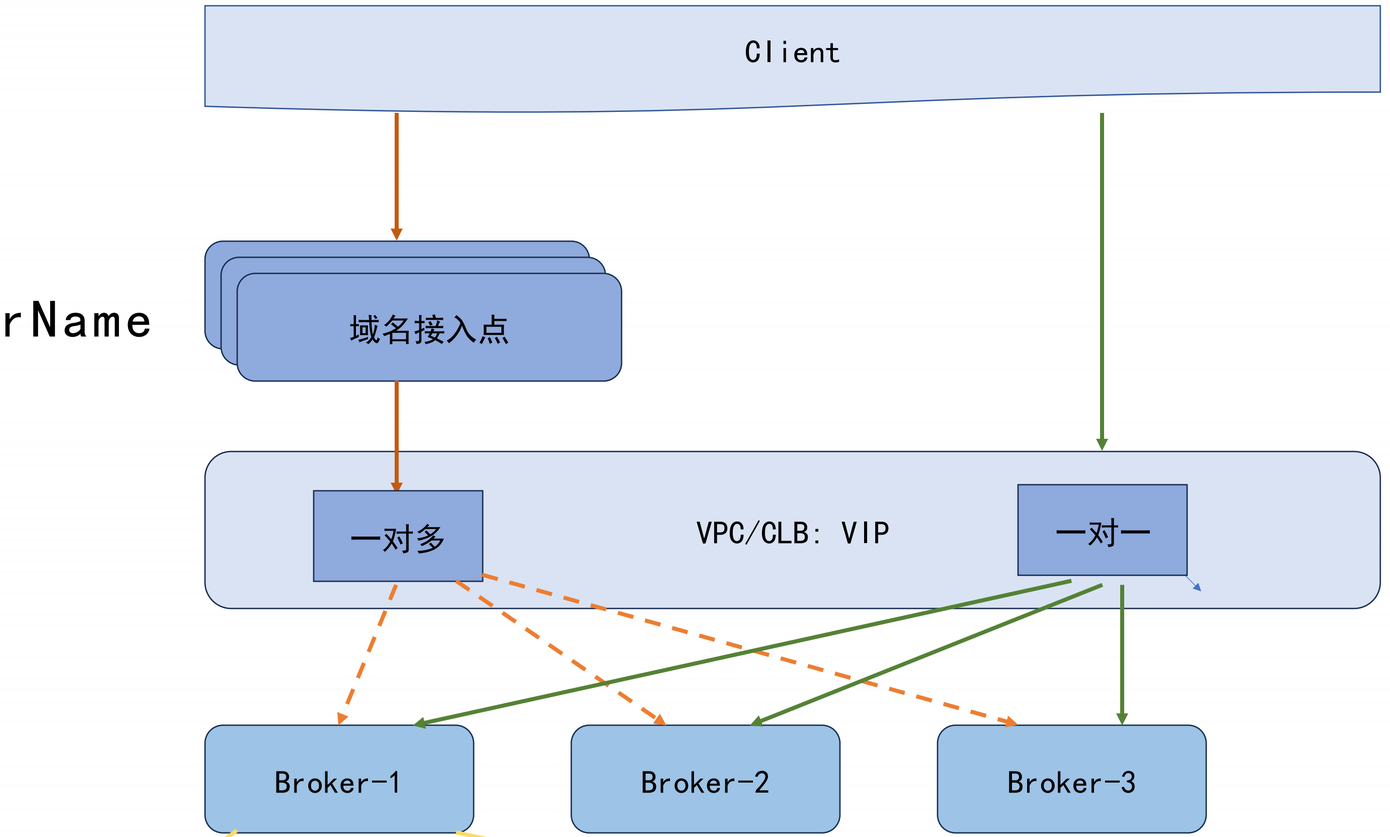
最初方案

AdvertisedListeners+ListenerName

1、配置复杂

2、职责混乱

3、维护成本高



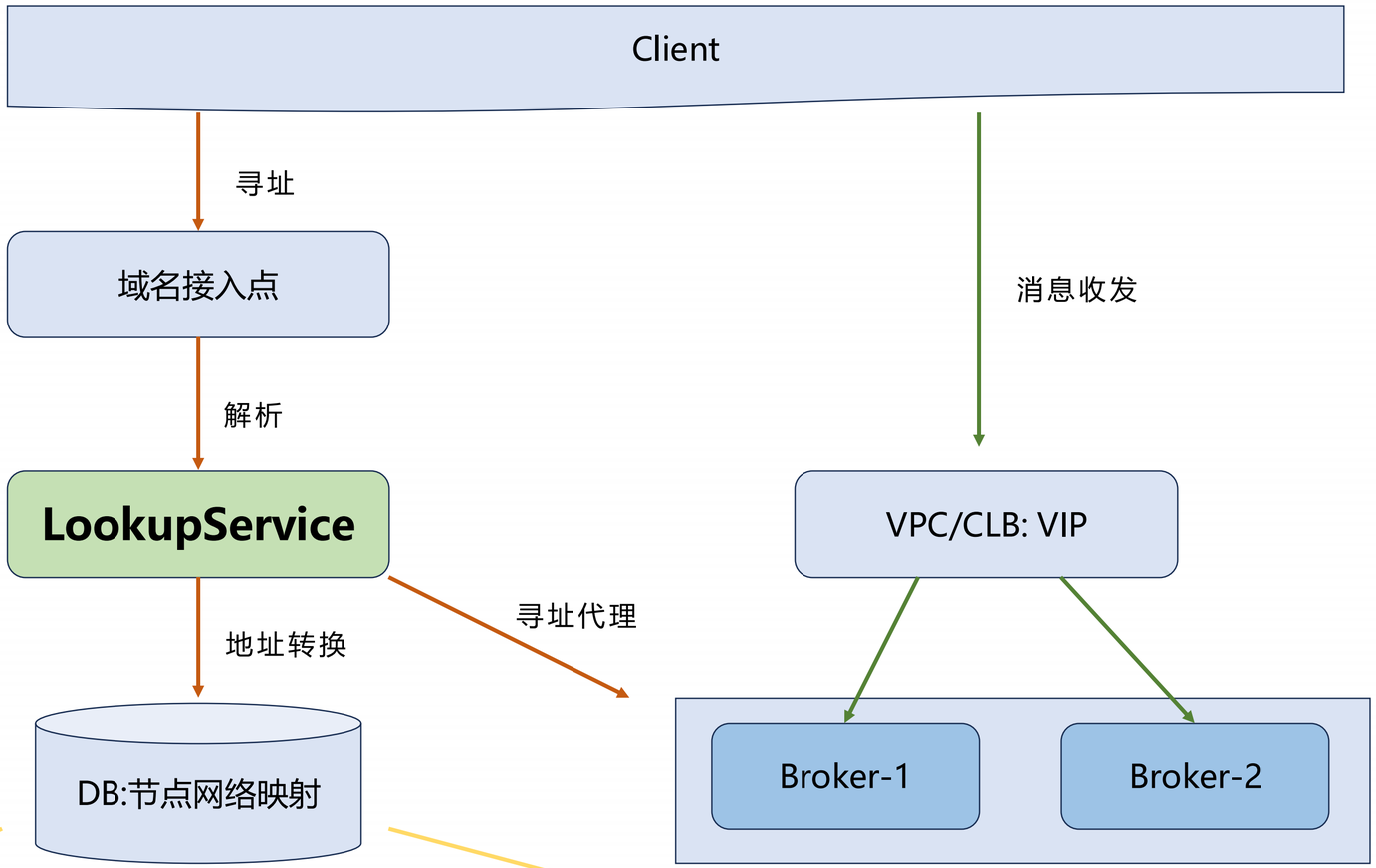

```
advertisedListeners=
INTERNAL:pulsar://10.1.0.1:6650, // 内网
PUBLIC:pulsar://43.198.0.1:10001, // 公网
VPC-1:pulsar://9.1.0.1:10001, // VPC-1
VPC-2:pulsar://9.2.0.1:10001 // VPC-2
... // 其他云私有网络等
```

1. 多网接入-路由寻址

改进方案

引入LookupService

1、简化架构
2、职责清晰
3、运维简单
4、扩展性强



节点	网络类型	映射地址
broker-1	INTERNAL(内网)	10.1.0.1:6650
broker-1	PUBLIC(公网)	43.198.0.1:10001
broker-1	VPC-1(私有网络)	9.1.0.1:10001
broker-1	VPC-2(私有网络)	9.2.0.1:10001

...

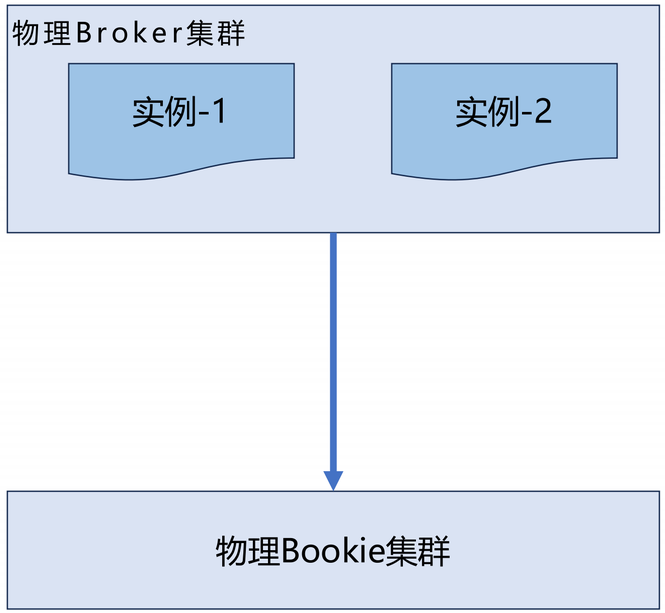
1.多网接入

2.集群迁移

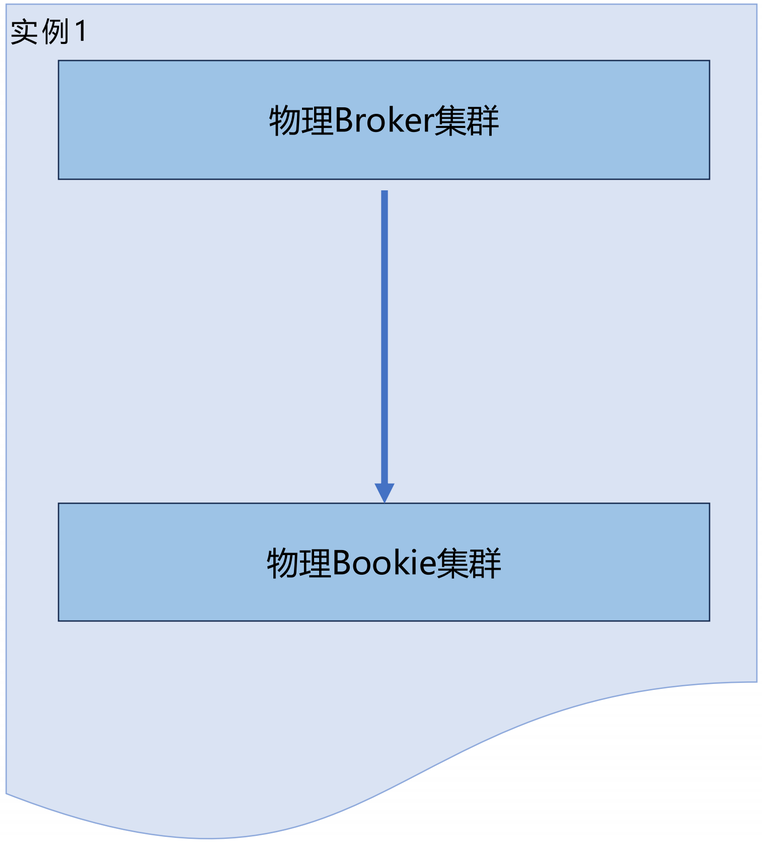
3.高可用最佳实践

2. 集群迁移–产品形态

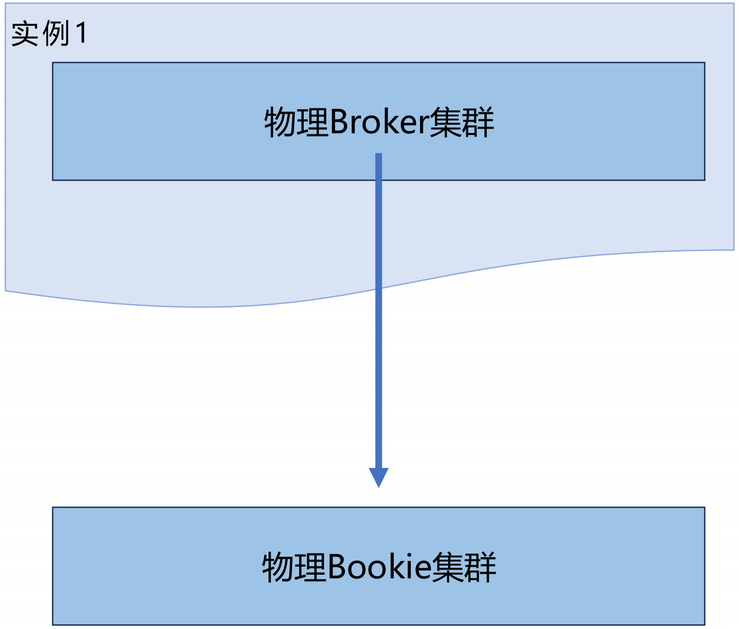
Broker共享，Bookie共享



Broker独占，Bookie独占



Broker独占，Bookie共享



2. 集群迁移–整体架构

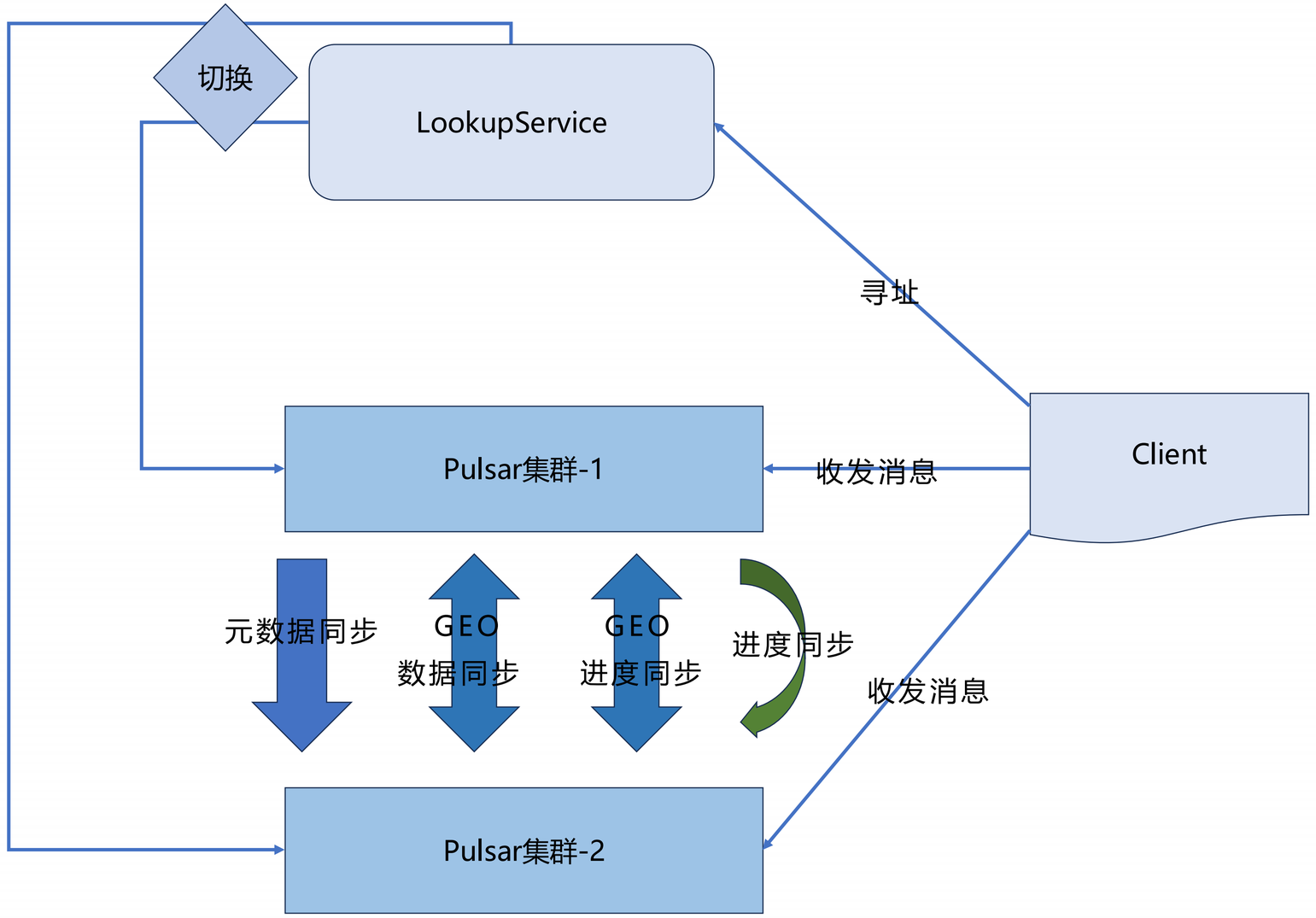
主要流程

1、元数据同步

2、数据同步（GEO）

3、订阅进度同步（GEO+补偿）

4、切换集群（Unload+寻址调整）



Tenant-a 从Pulsar集群-1迁移到Pulsar集群-2

2. 集群迁移-订阅进度说明

订阅进度：

MarkDeletePosition = 1:2

IndividualDeleteMessages = [(1:3-1:4], (1:6-1:7]]

消息id：	1:8	1:7	1:6	1:5	1:4	1:3	1:2	1:1	1:0
消费状态：	未确认	未确认	已确认	未确认（延迟消息）	已确认	未确认	已确认	已确认	已确认

GEO只同步MarkDeletePosition

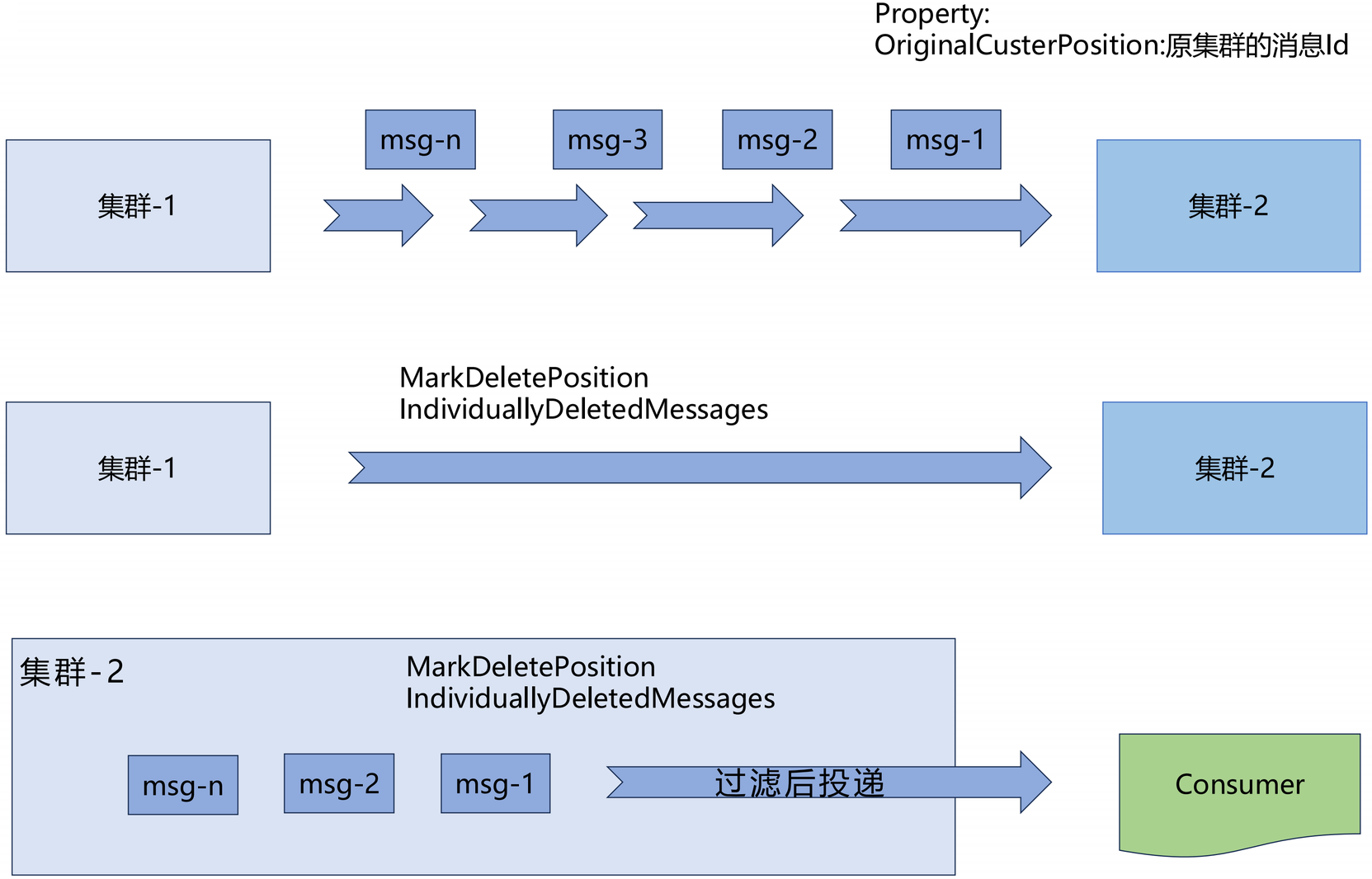
且很多情况下也不能保证同步成功

2. 集群迁移-进度同步

1、携带原集群消息id

2、原集群进度同步

3、消费阶段过滤

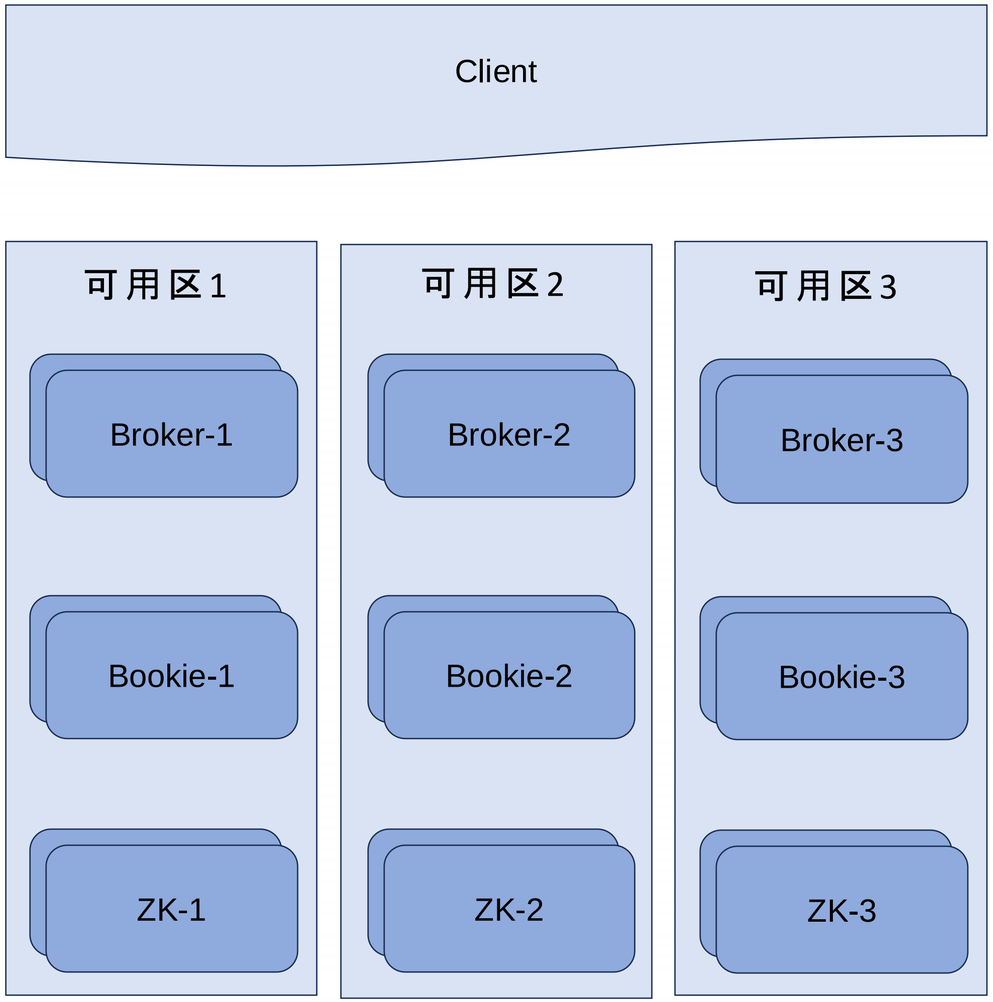


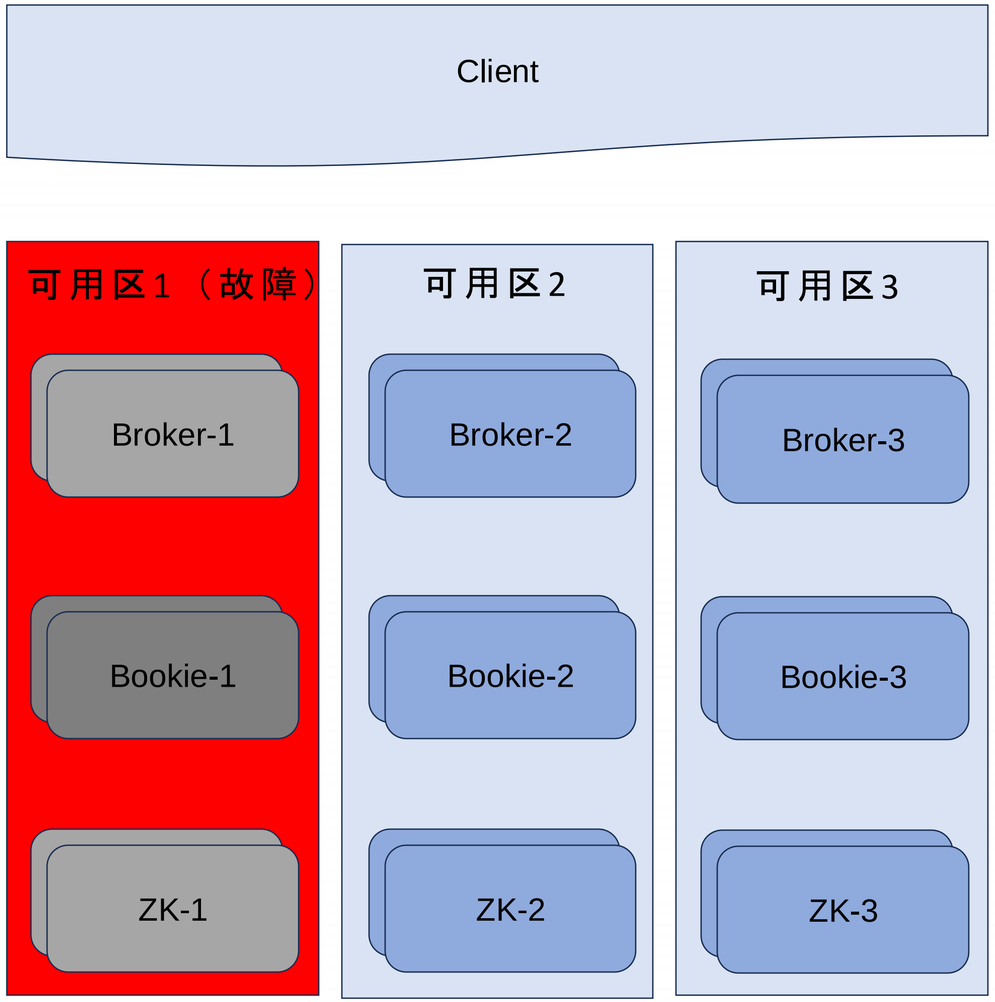
1.多网接入

2.集群迁移

3.高可用最佳实践

3. 高可用-可用区容灾


Client
可用区1
Broker-1
Bookie-1
ZK-1
可用区2
Broker-2
Bookie-2
ZK-2
可用区3
Broker-3
Bookie-3
ZK-3


Client
可用区1（故障）
Broker-1
Bookie-1
ZK-1
可用区2
Broker-2
Bookie-2
ZK-2
可用区3
Broker-3
Bookie-3
ZK-3

3. 高可用-副本分布

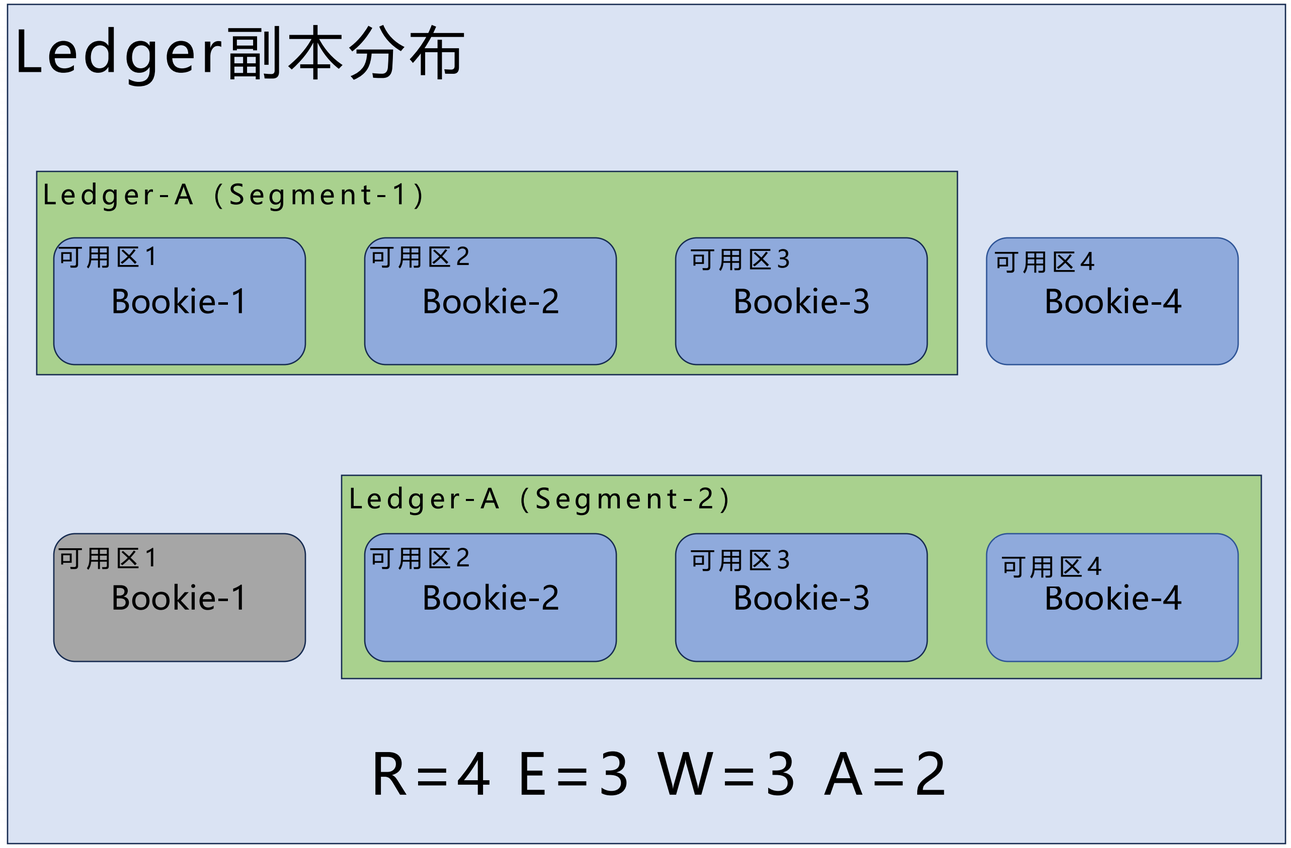
1、机架感知

2、跨可用区分布

配置:
bookkeeperClientRackawarePolicyEnabled=false
bookkeeperClientMinNumRacksPerWriteQuorum=3

zk节点:

```
{
    "default": {
        "Bookie-1:3181": {
            "rack": "zone-1"
        },
        "Bookie-2:3181": {
            "rack": "zone-2"
        },
        "Bookie-3:3181": {
            "rack": "zone-3"
        },
        "Bookie-4:3181": {
            "rack": "zone-4"
        }
    }
}
```

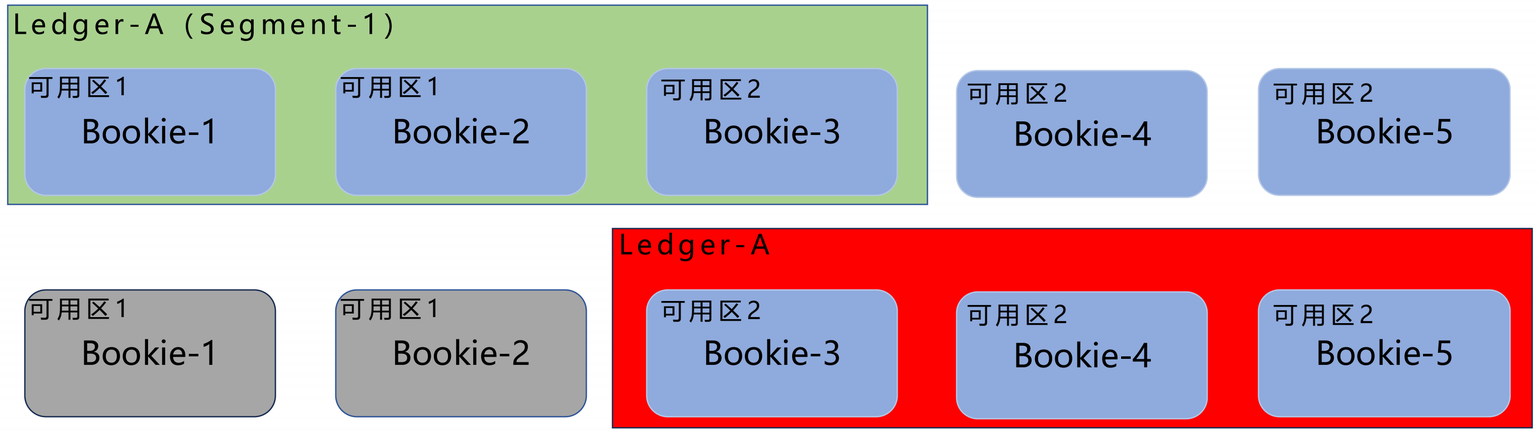


3. 高可用-最佳实践

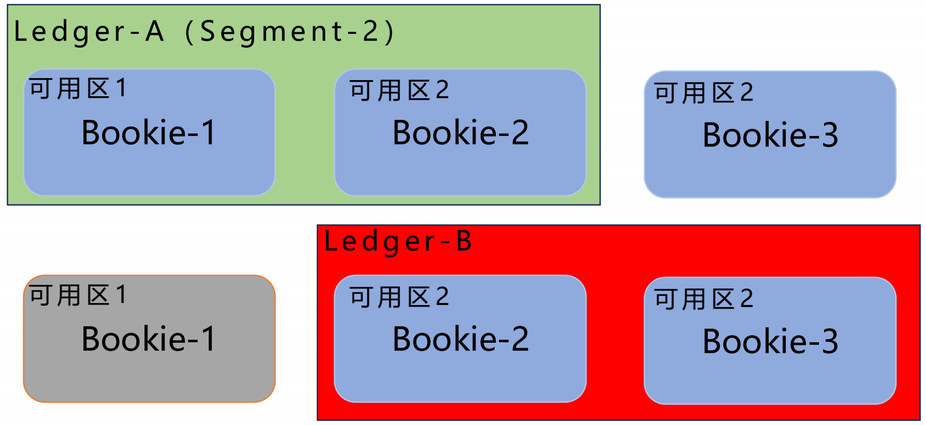
恢复过程

读取LAC（Last-Add-Confirm）

剩余副本 >= W - A + 1



R=2 E=3 W=3 A=2 无法恢复



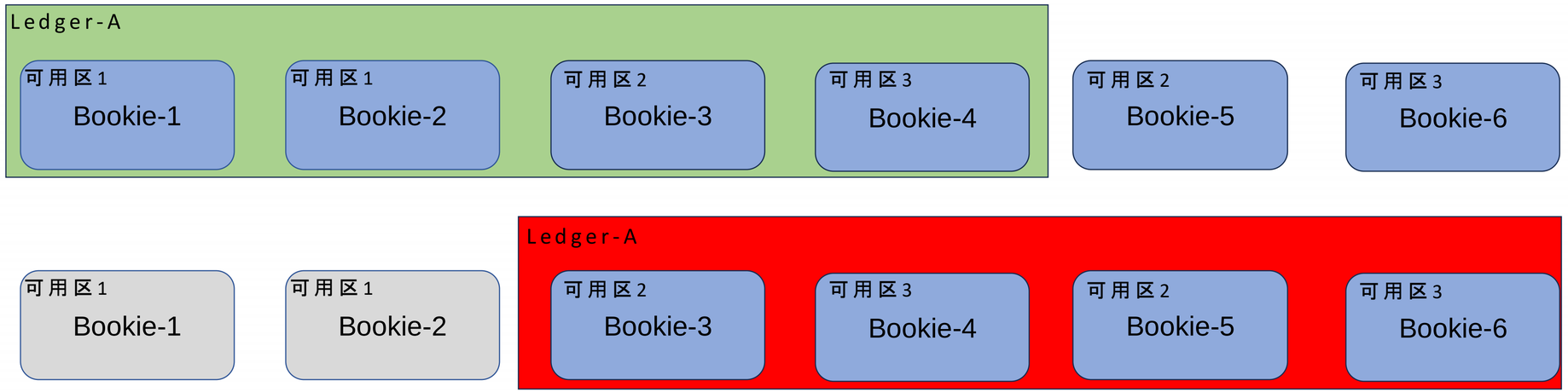
R=2 E=2 W=2 A=1 **无法恢复**

3. 高可用-最佳实践

恢复过程

读取LAC（Last-Add-Confirm）

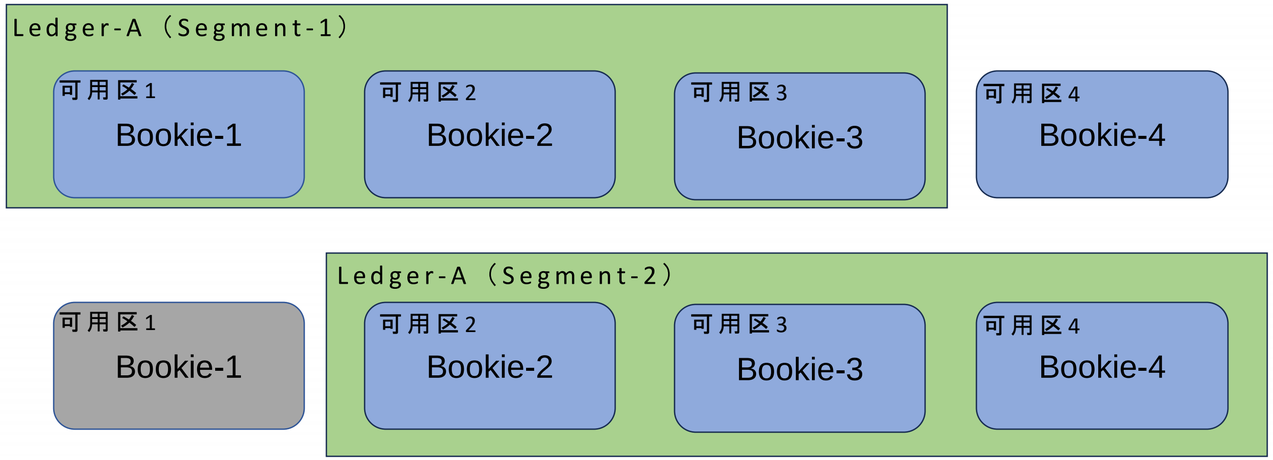
剩余副本 >= W - A + 1



R=3 E=4 W=3 A=2 无法恢复

3. 高可用-最佳实践


```
"allEnsembles" : {
  "0" : [ {
    "id" : "10.1.0.1:3181"
  }, {
    "id" : "10.1.0.2:3181"
  }, {
    "id" : "10.1.0.3:3181"
  } ]
  "10000" :  [ {
    "id" : "10.1.0.4:3181"
  }, {
    "id" : "10.1.0.2:3181"
  }, {
    "id" : "10.1.0.3:3181"
  } ]
},
```



E=3 W=3 A=2

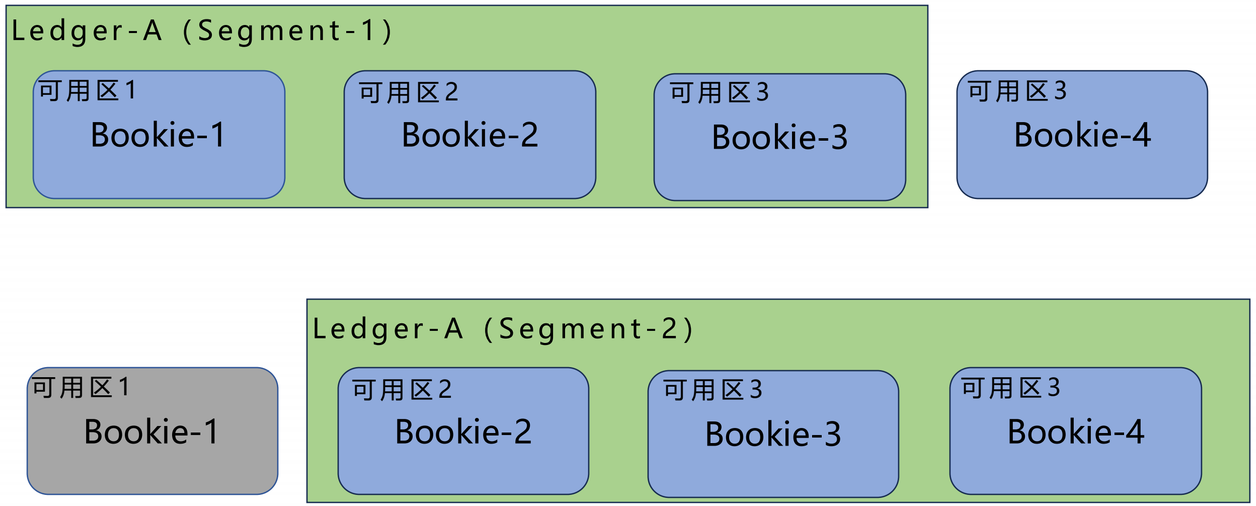
BookkeeperEnableStickyReads=true && E=W

3. 高可用-最佳实践

3可用区

E=3 W=3 A=2

bookkeeperEnableStickyReads=true

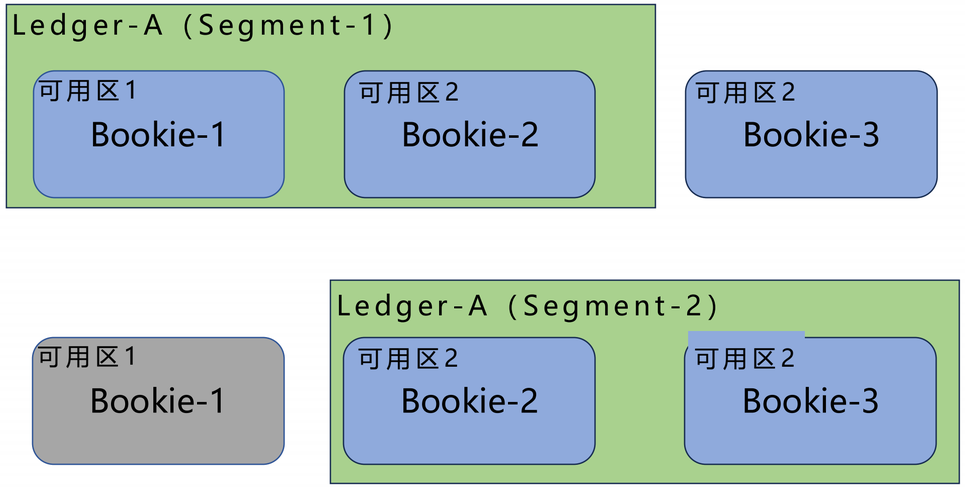


R=3 E=3 W=3 A=2

2可用区

E=2 W=2 A=2

bookkeeperEnableStickyReads=true



R=2 E=2 W=2 A=2

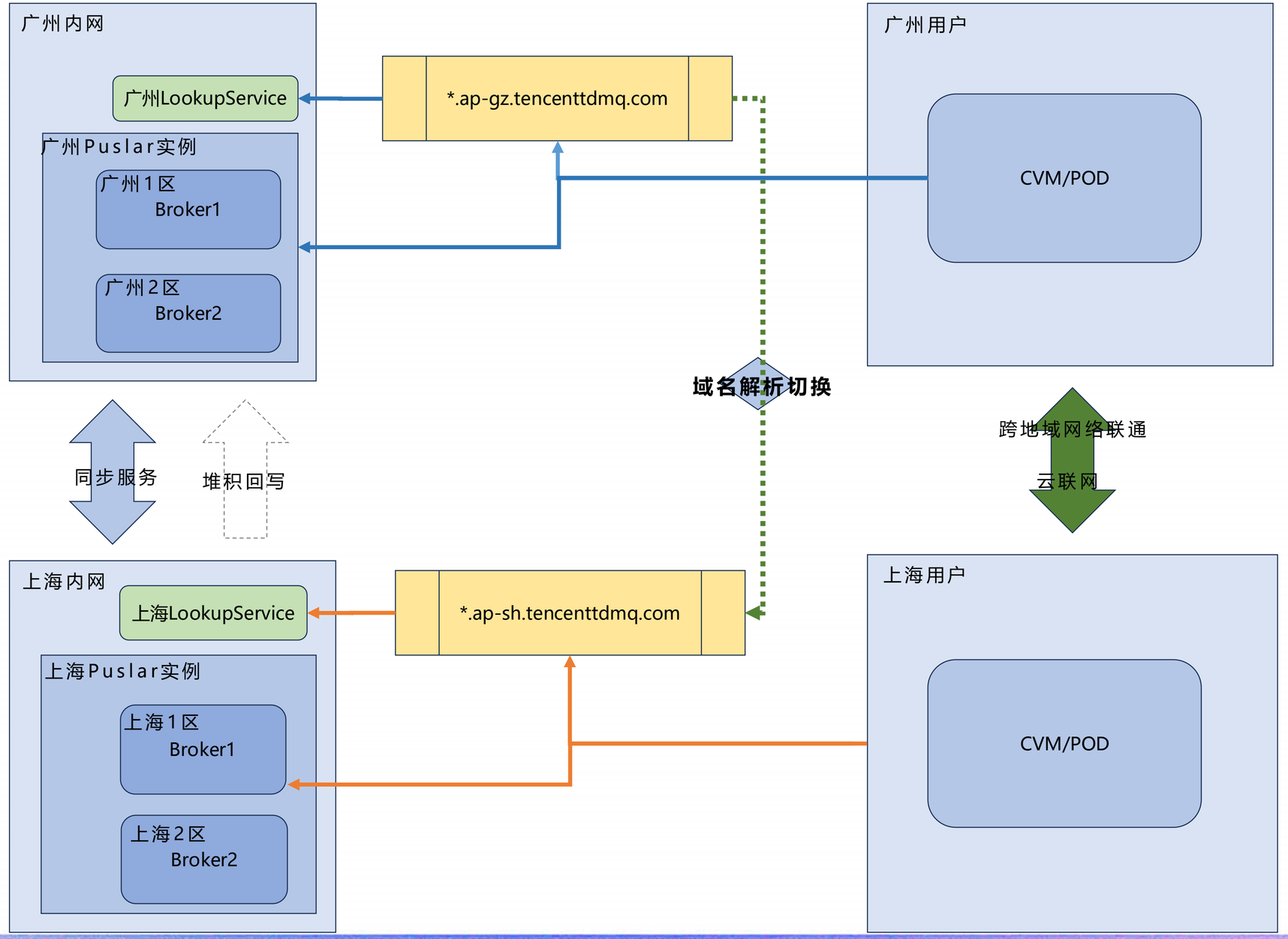
3. 高可用-地域容灾

1、异地备集群

2、元数据同步

3、域名切换

4、堆积消息回写



PULSAR Meetup | Thanks

韩明泽 腾讯
hanmzarsenal@gmail.com
王震江 腾讯
zhenjiang_wang@qq.com