系统架构——分层 + 分片

演讲者 / streamlio 翟佳

资料整理
火龙果软件
荐

Apache Pulsar特性

Durability
Data replicated and synced to disk

Ordering
Guaranteed ordering

Delivery Guarantees
At least once, at most once and effectively once

High throughput
Can reach 1.8 M messages/s in a single partition

Low Latency
Low publish latency of 5ms at 99pct

Unified messaging model
Support both Streaming and Queuing

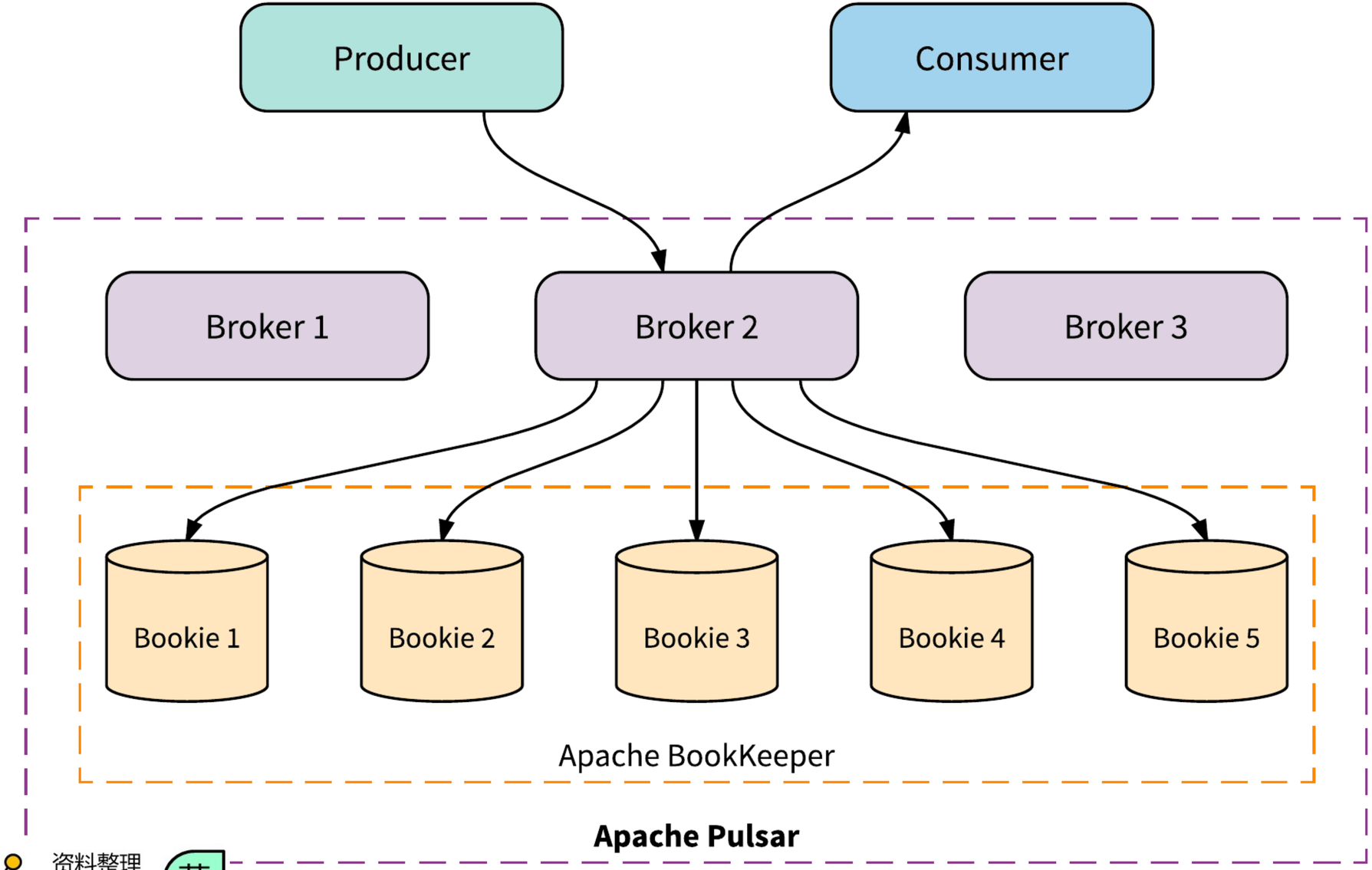
Multi-tenancy
A single cluster can support many tenants and use cases

Geo-replication
Out of box support for geographically distributed applications

Highly scalable & available
Can support millions of topics

资料整理
火龙果软件
荐

分层架构

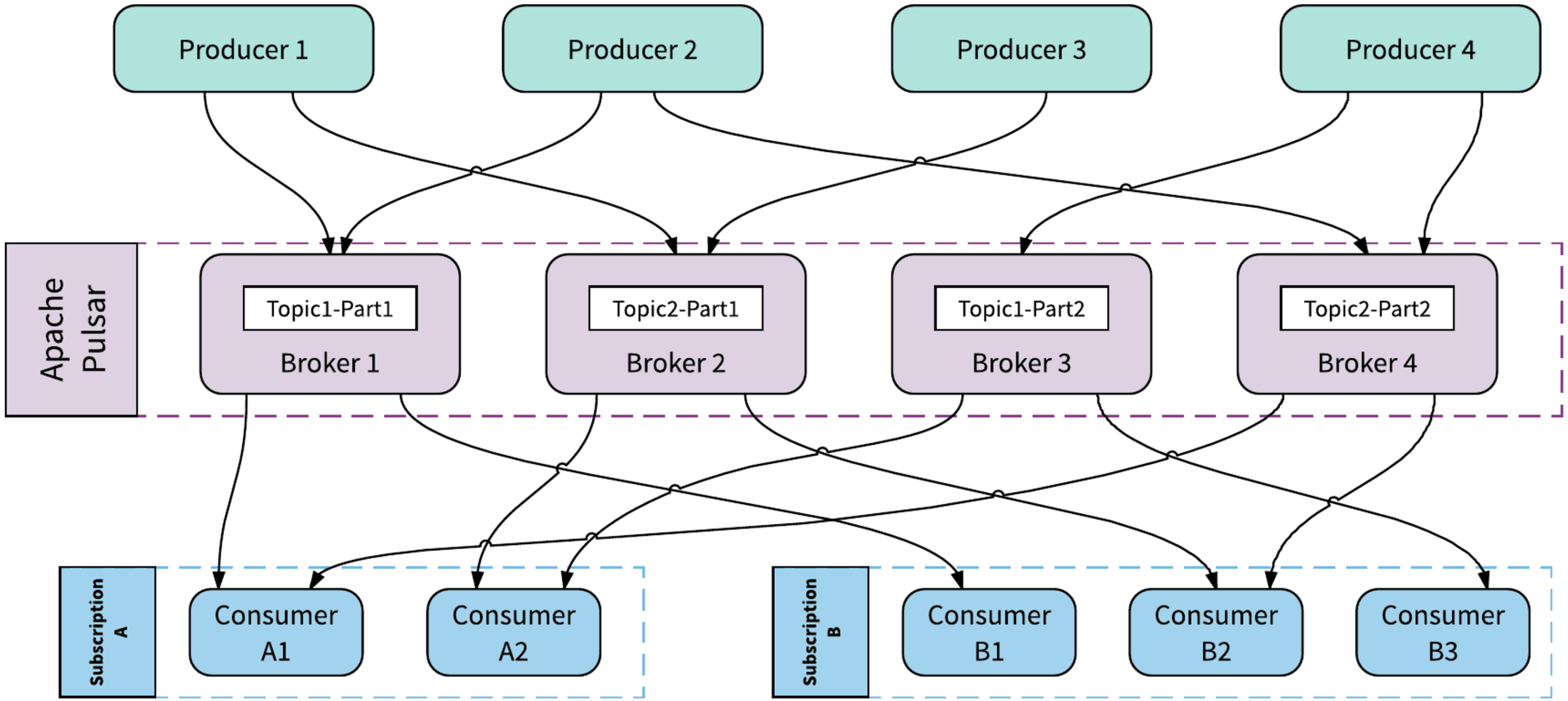


分层架构:
Brokers & Bookies

- 独立扩展
- 灵活容错
- 快速扩容

资料整理
火龙果软件
荐

分层架构 — Brokers

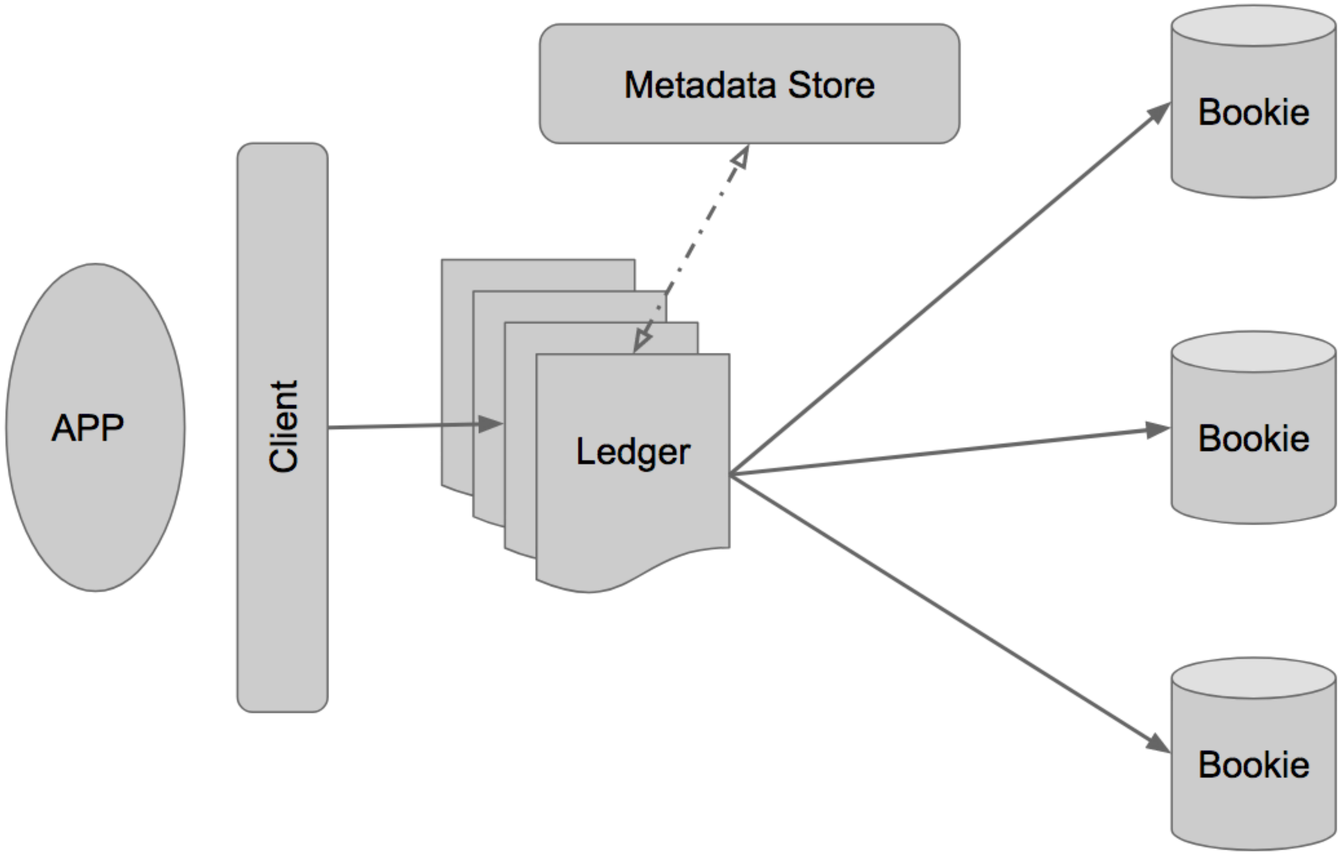


资料整理
火龙果软件
荐

分层架构 — BookKeeper

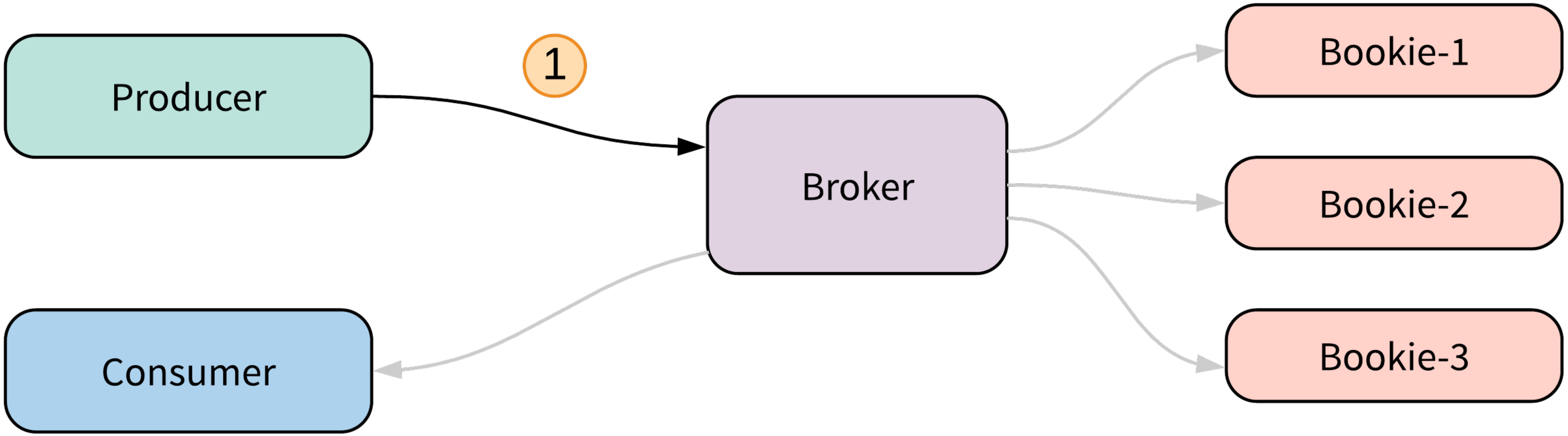
BookKeeper分布式日志/流存储

- 低延时、高吞吐、持久化
- 强一致 (repeatable read consistency)
- 高可用
- 单节点可以存储很多日志
- I/O隔离



资料整理
火龙果软件
荐

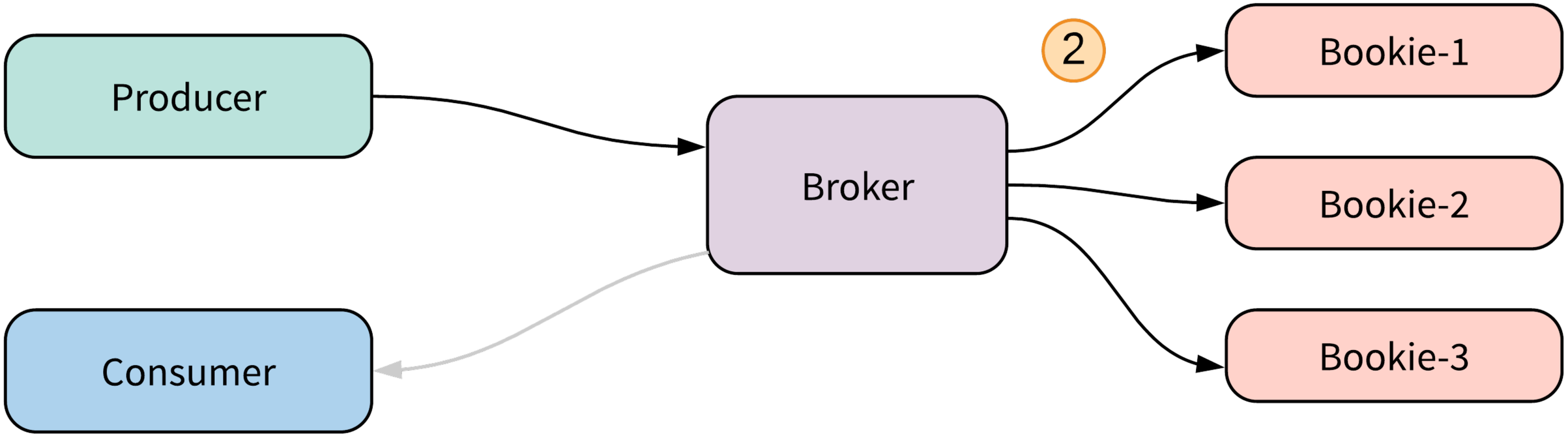
读写流程 (1)



(1) 生产者发送消息给Broker

资料整理
火龙果软件
荐

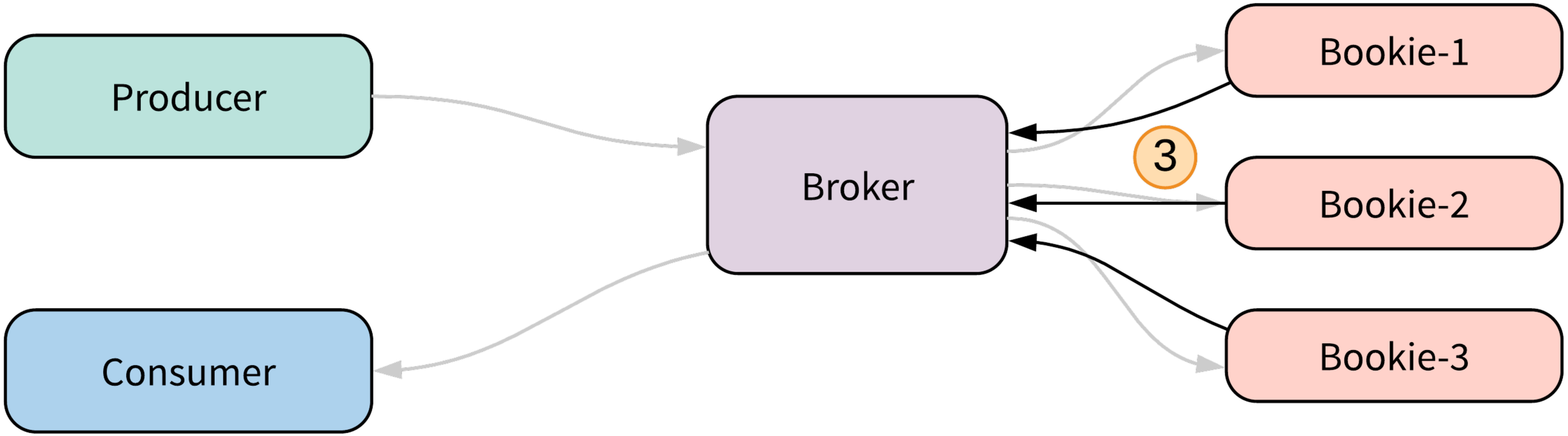
读写流程 (2)



(2) Broker并发写N个副本

资料整理
火龙果软件
荐

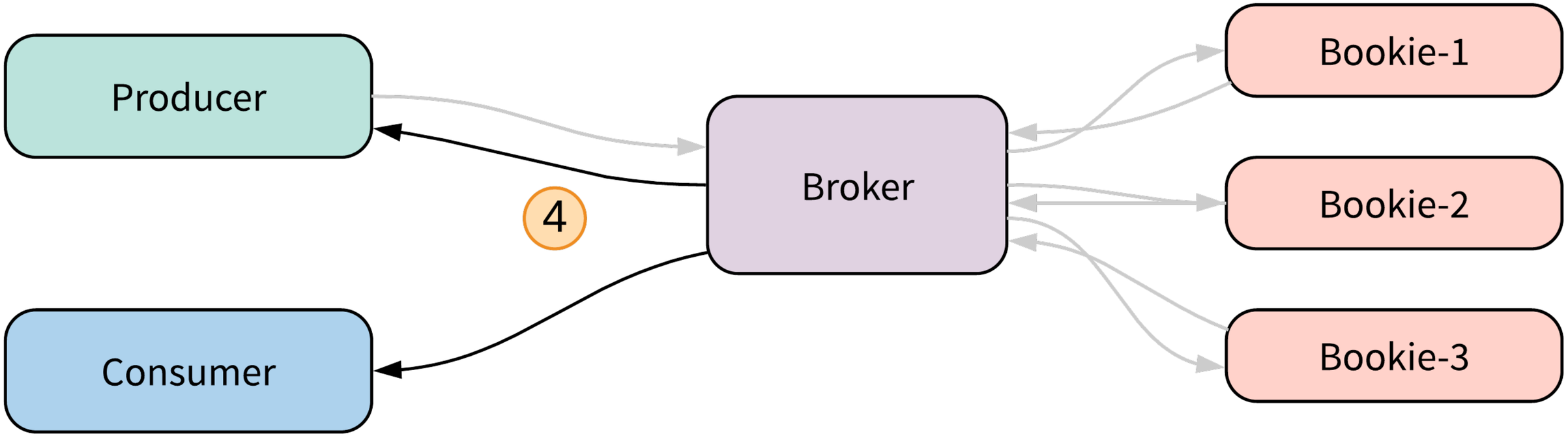
读写流程 (3)



(3) Broker等待来自于Bookies的Quorum Acks

资料整理
火龙果软件
荐

读写流程 (4)

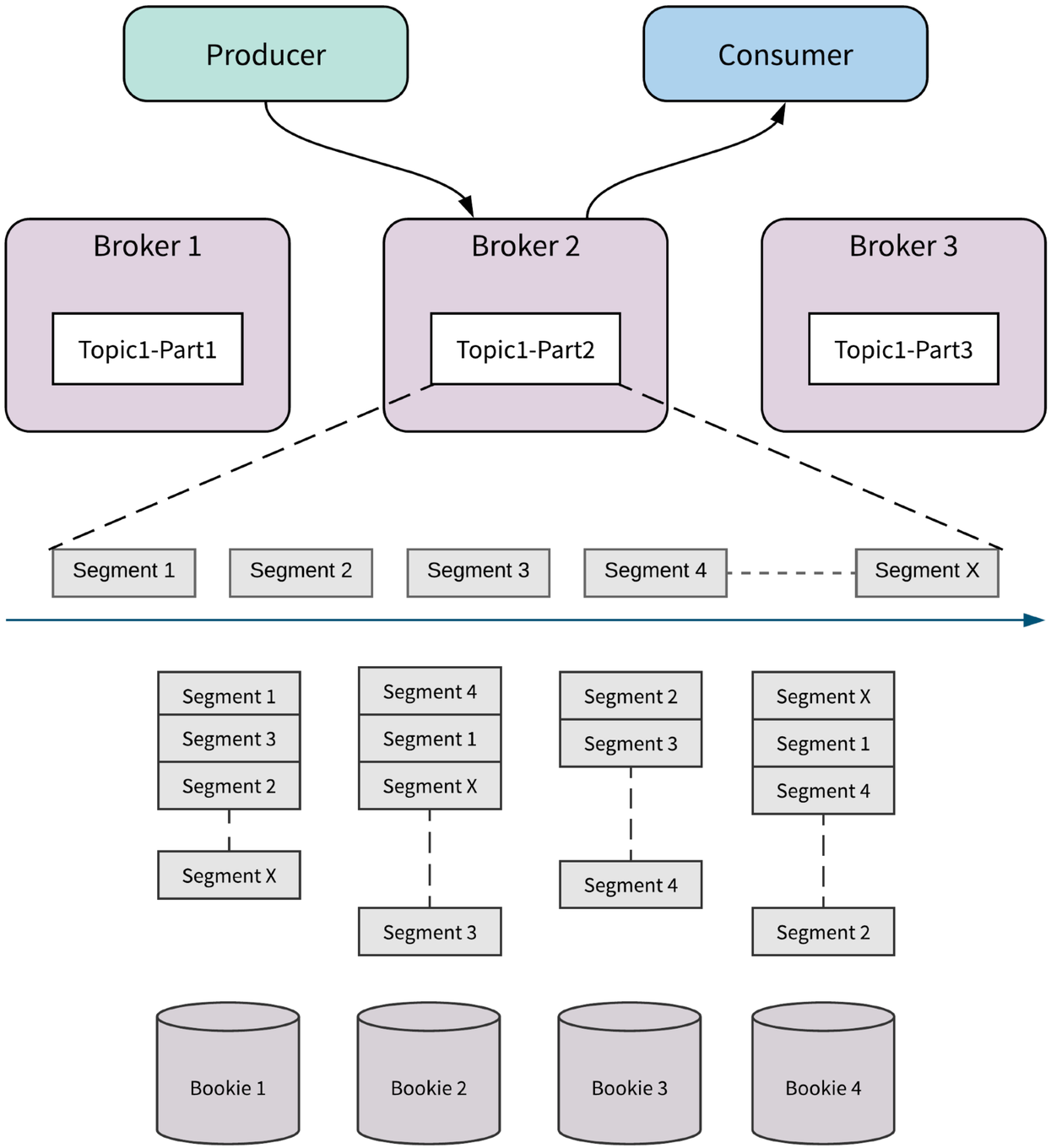


(4) Broker返回确认给生产者，并投递给消费者

资料整理
火龙果软件
荐

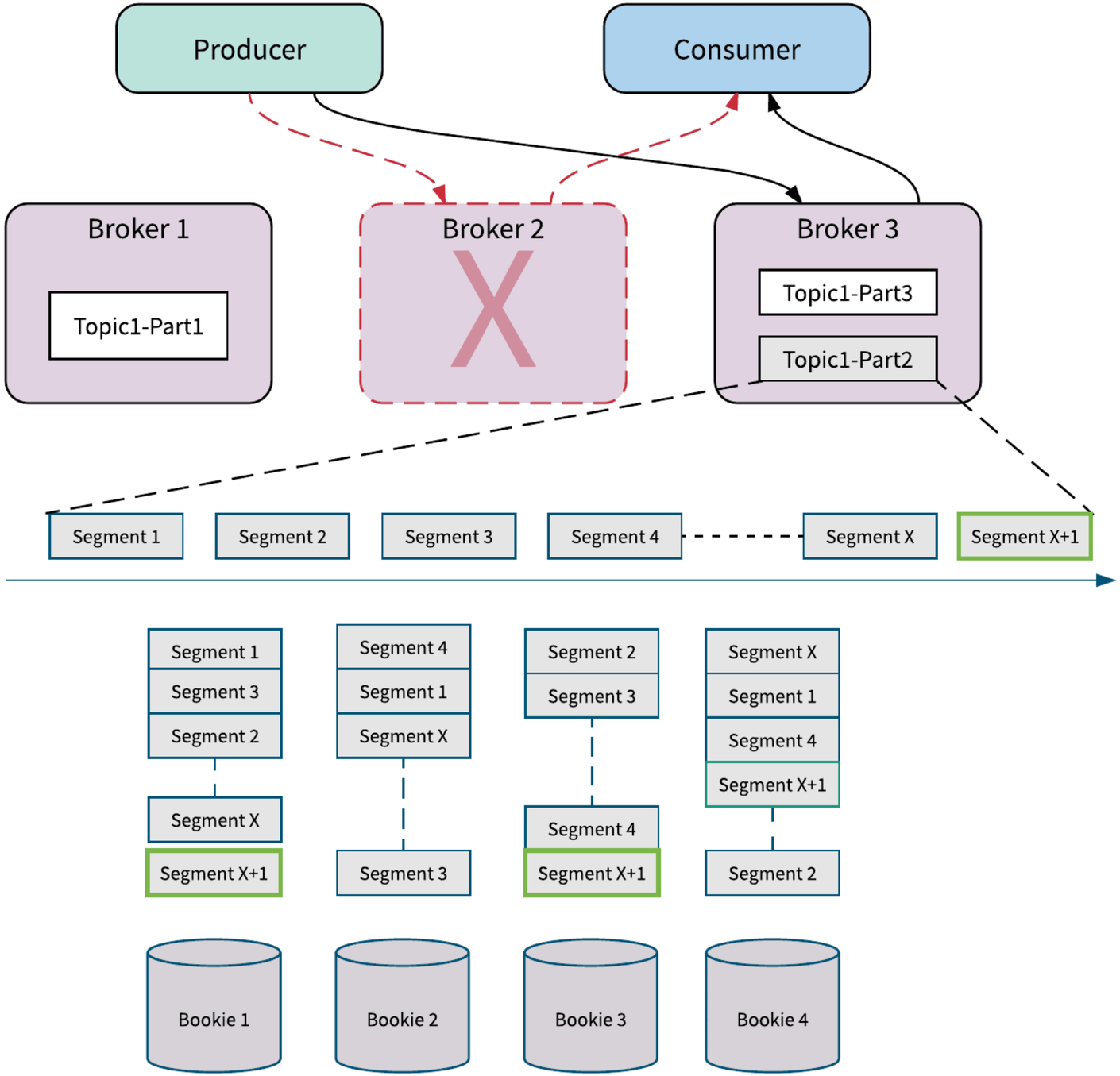
分片存储

- 分区(Partition)是逻辑上的一个概念
- 分区按照时间或者大小被切成分片(Segment)
- 分片被打散存放到集群中的所有节点



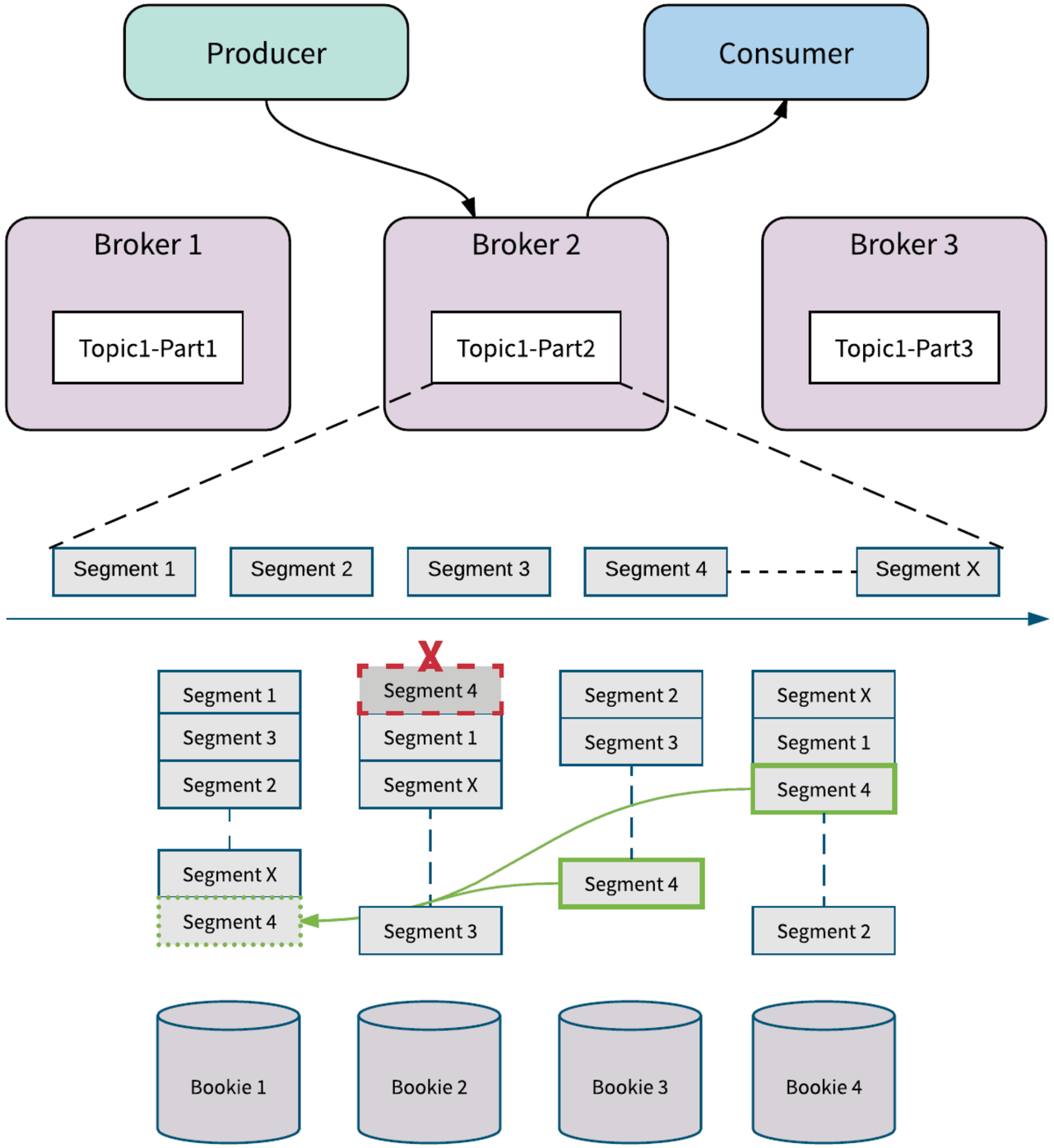
资料整理
火龙果软件
荐

Broker容错



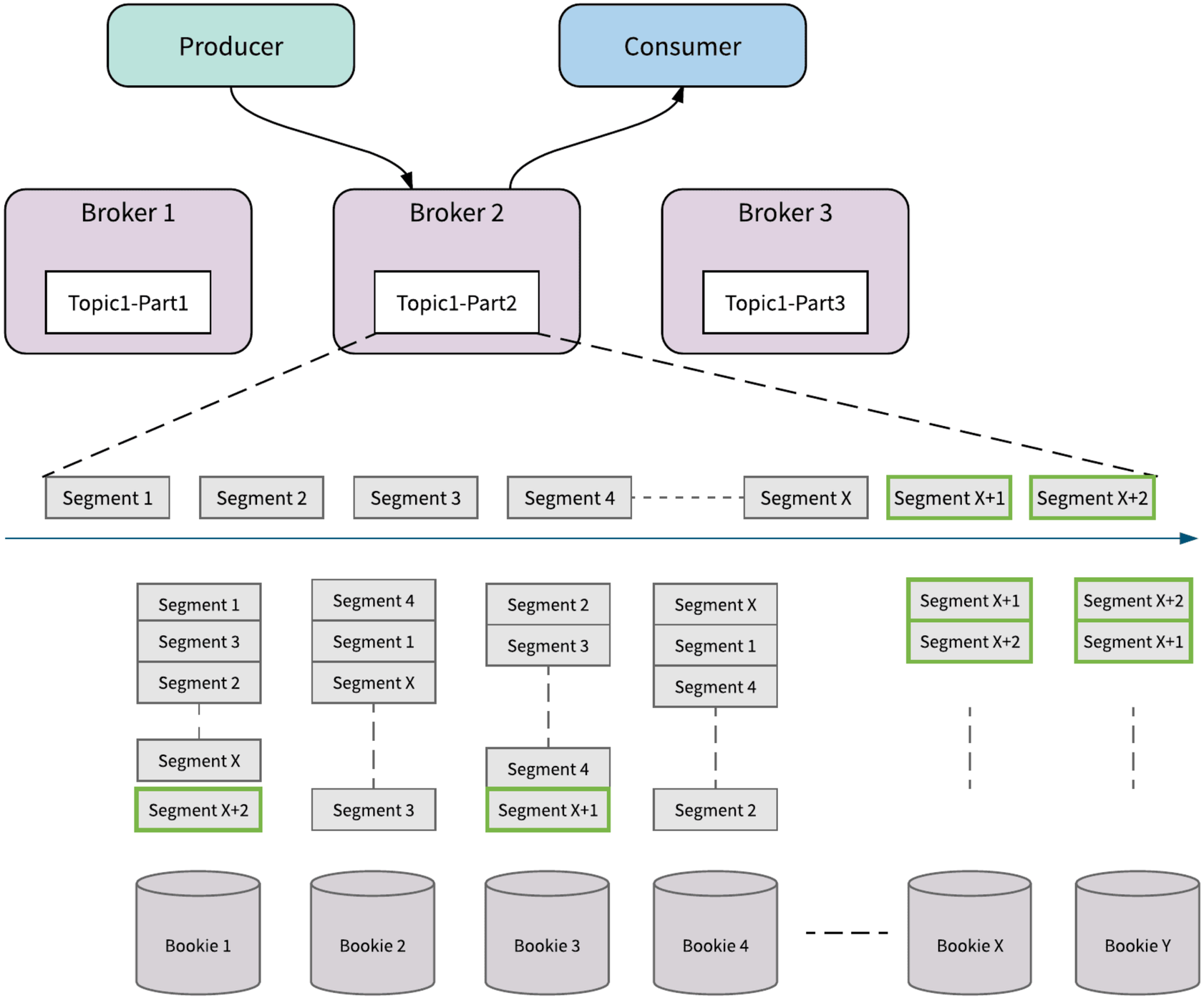
资料整理
火龙果软件
荐

Bookie容错



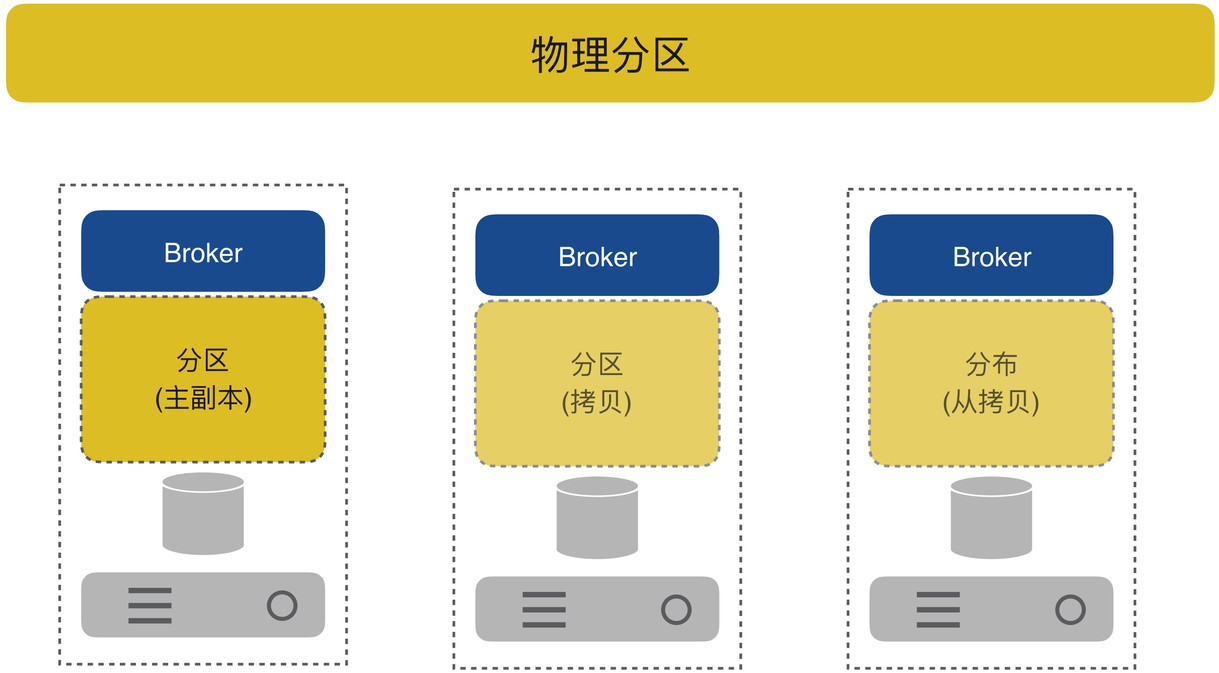
资料整理
火龙果软件
荐

扩容



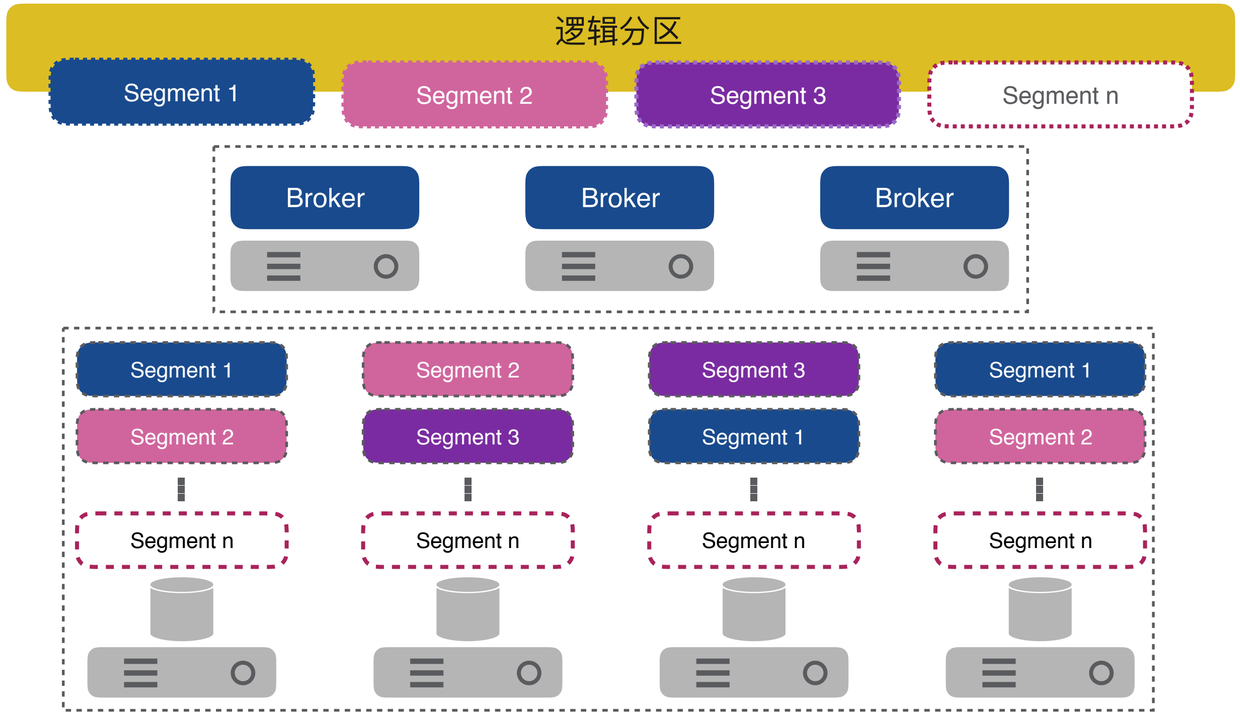
资料整理
火龙果软件
荐

分区 vs 分片



分区架构

- 物理分区
- 存储和计算紧耦合
- 容错恢复需要拷贝物理分区
- 扩容需要迁移物理分区来达到负载均衡

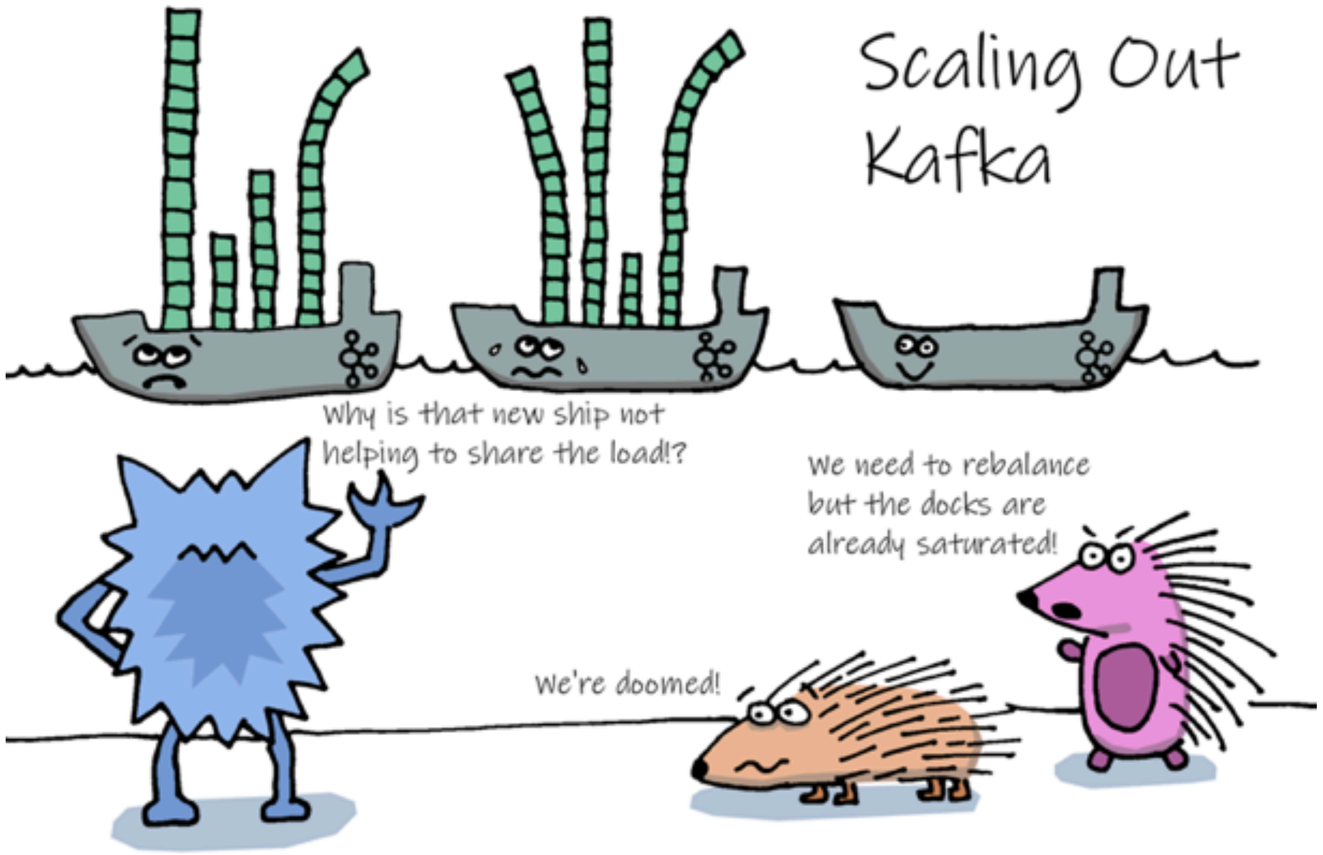


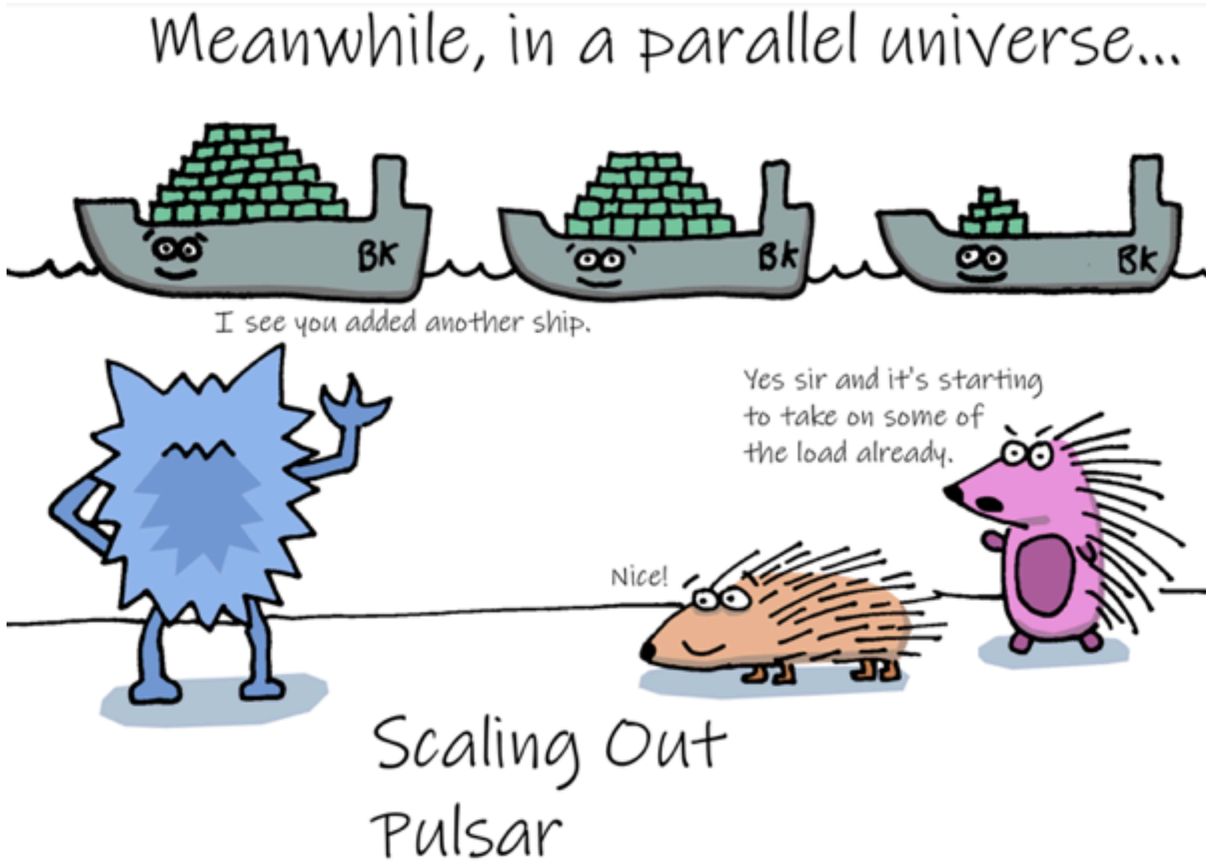
分片架构

- 逻辑分区，“物理”分片
- 存储和计算分离
- 失效处理相互分离，快速、无痛点
- 弹性扩容

资料整理
火龙果软件
荐

分区 vs 分片





资料整理
火龙果软件
荐

lightly.com/sketches/2018/10/2/kafka-vs-pulsar-rebalancing-sketch

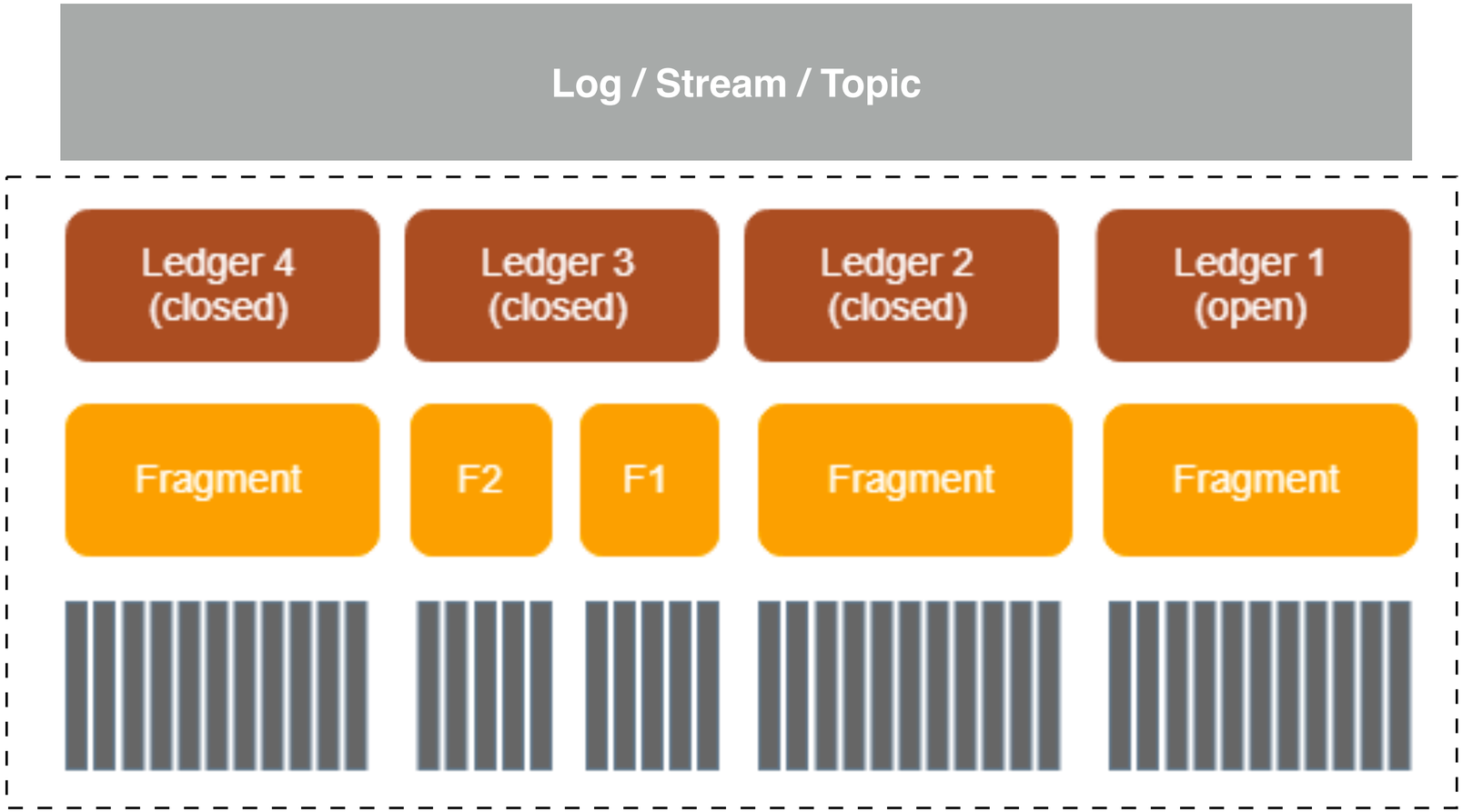
Apache BookKeeper

分布式日志/流存储

- 低延时多复本复制:
  - Quorum Parallel Replication
- 持久化: 所有操作保证刷盘后才进行ACK
- 强一致性:
  - 可重复读的一致性 (Repeatable Read Consistency)
- 读写高可用
- 存储节点的读写隔离

资料整理
火龙果软件
荐

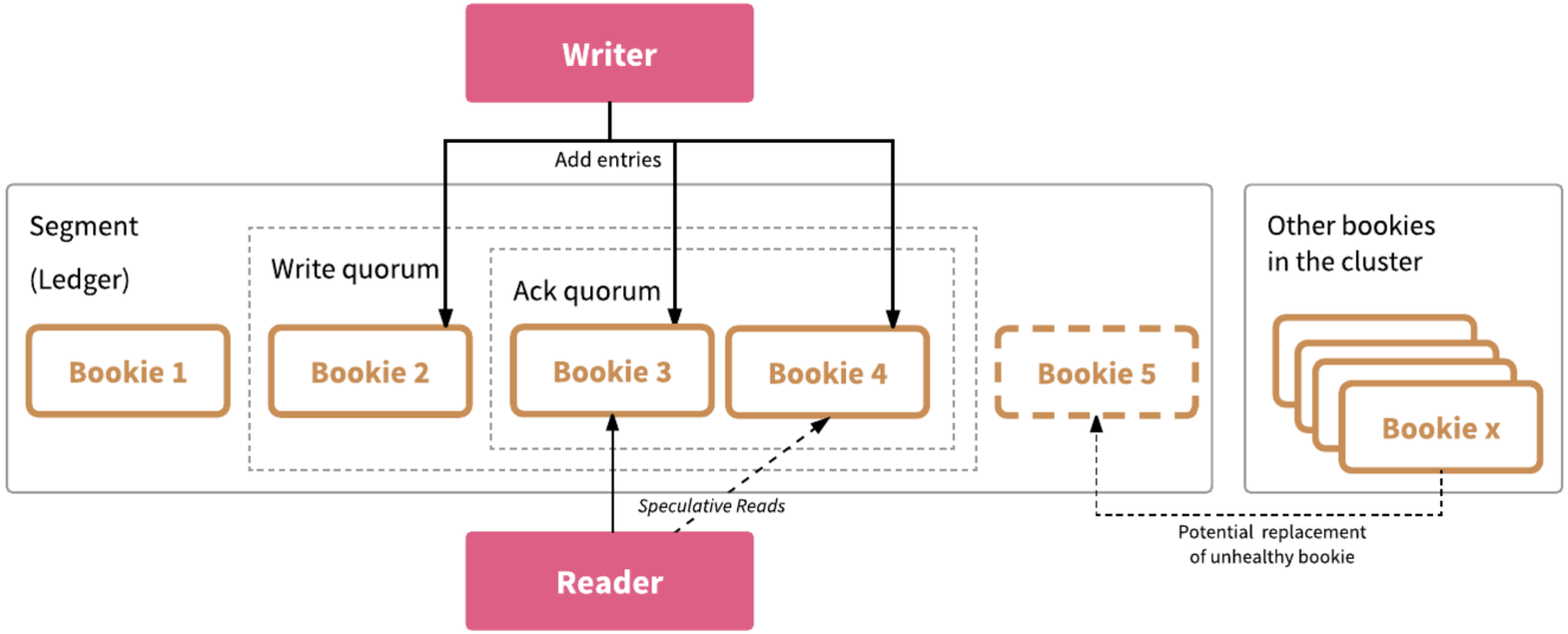
概念： Entry, Ledger & Log



资料整理
火龙果软件
荐

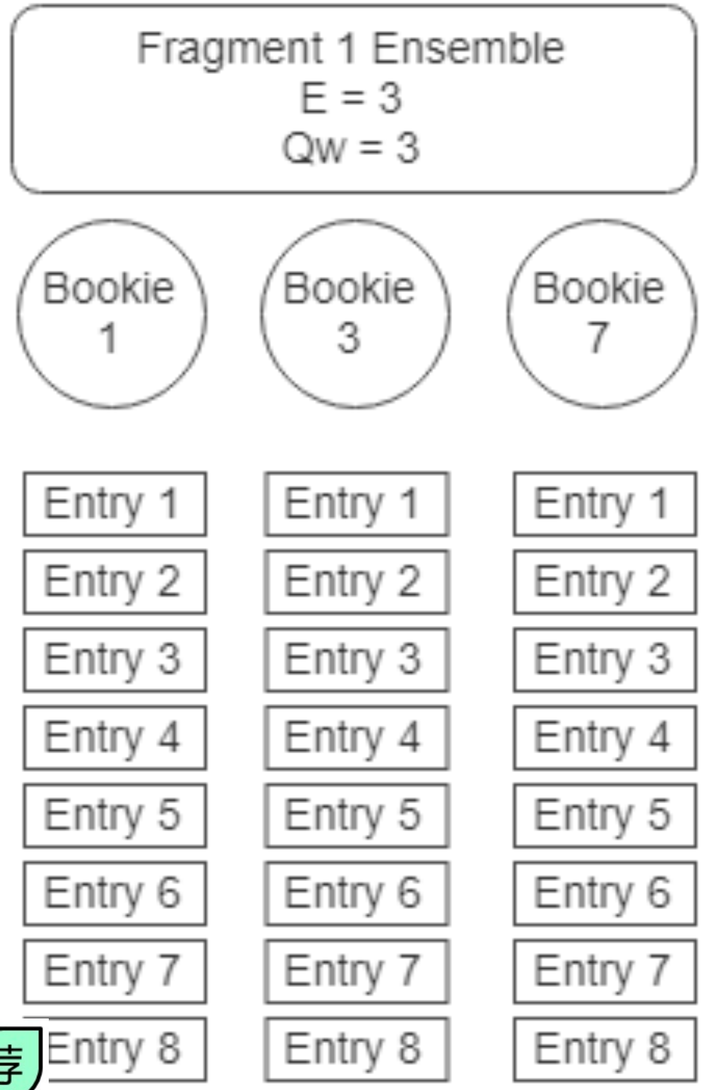
多副本复制(1)

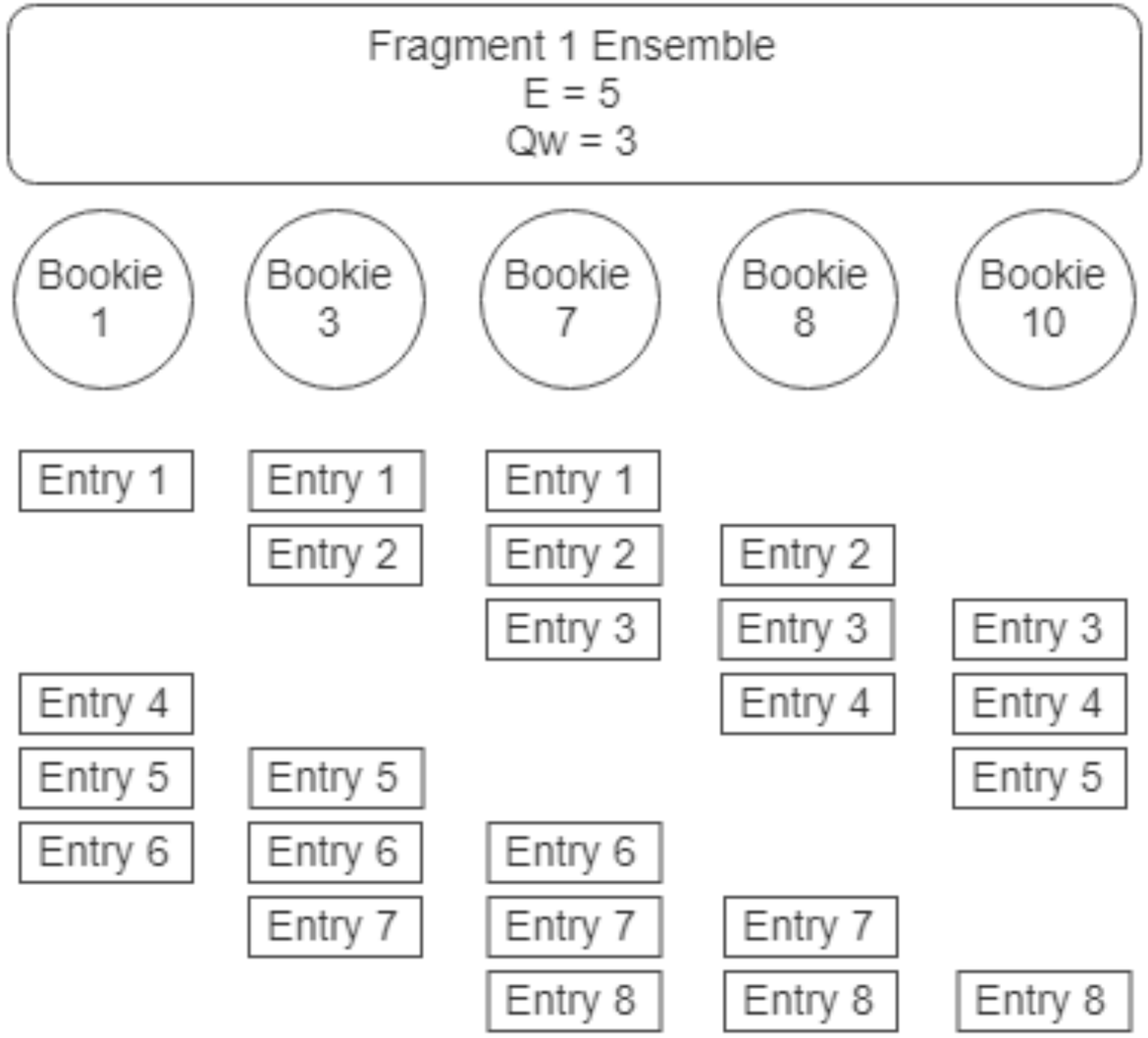
- Parallel Quorum Replication
- Ensemble
- Write Quorum
- Ack Quorum



资料整理
火龙果软件
荐

多副本复制(2)





资料整理
火龙果软件
荐

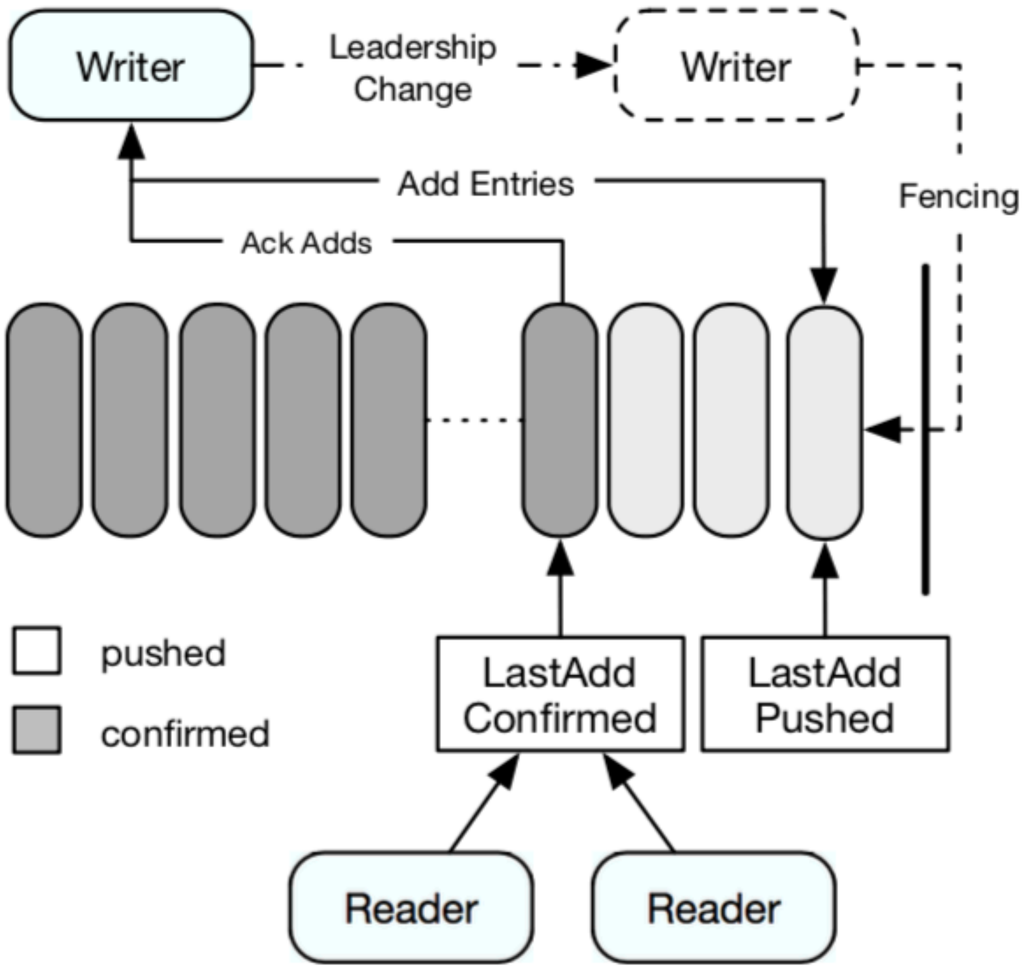
多副本复制(3)

- Ensemble: 控制一个Ledger的读写带宽
- Write Quorum：控制一条记录的复本数量
- Ack Quorum：写每条记录需要等待的Ack数量，控制时延
- 灵活性：
  - 增加Ensemble，可以增加读写带宽
  - 减少Ack Quorum，可以减少长尾时延

资料整理
火龙果软件
荐

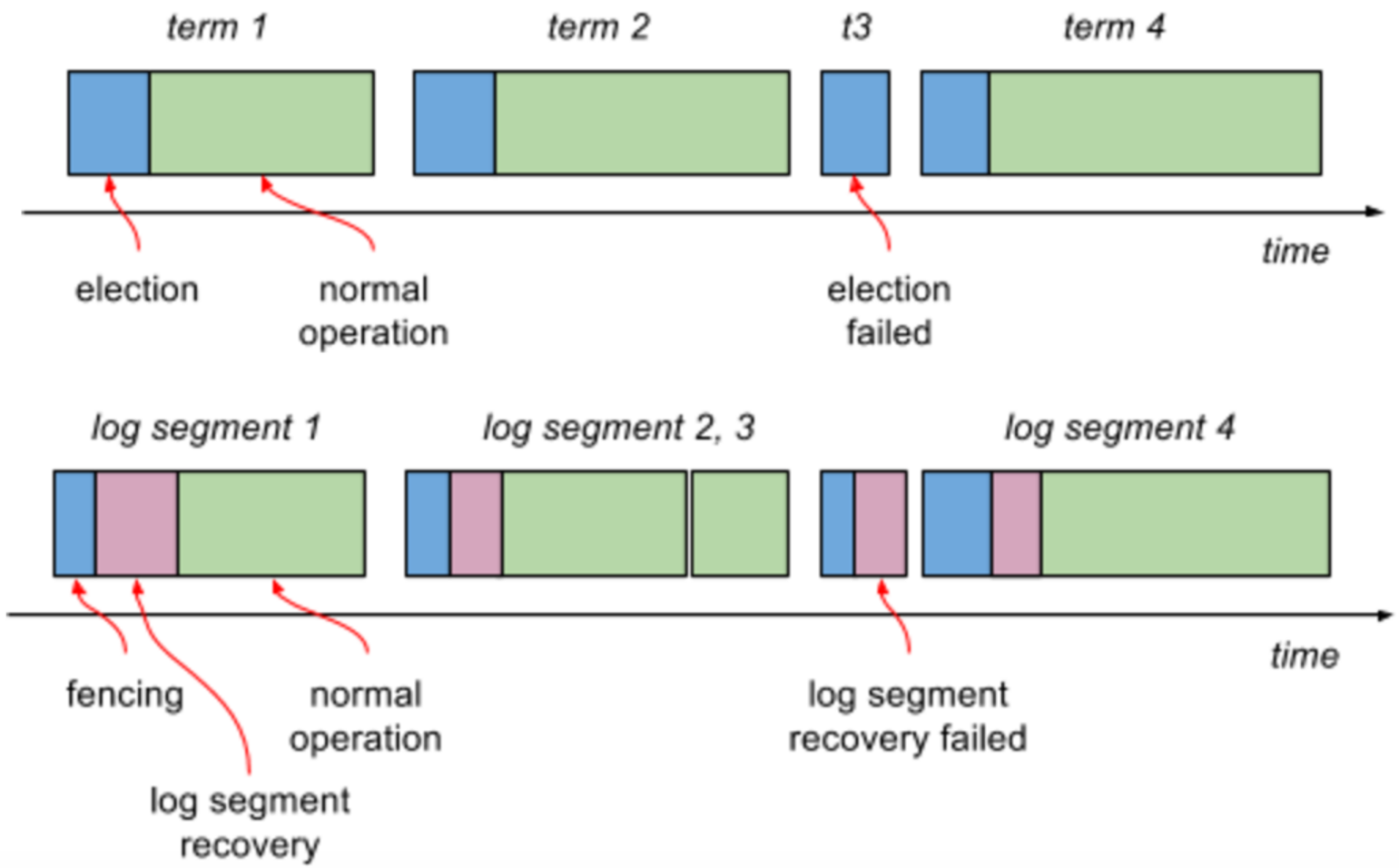
一致性(1)–可重复读

- LastAddPushed
- LastAddConfirmed
- Fencing避免脑裂



资料整理
火龙果软件
荐

一致性(2)

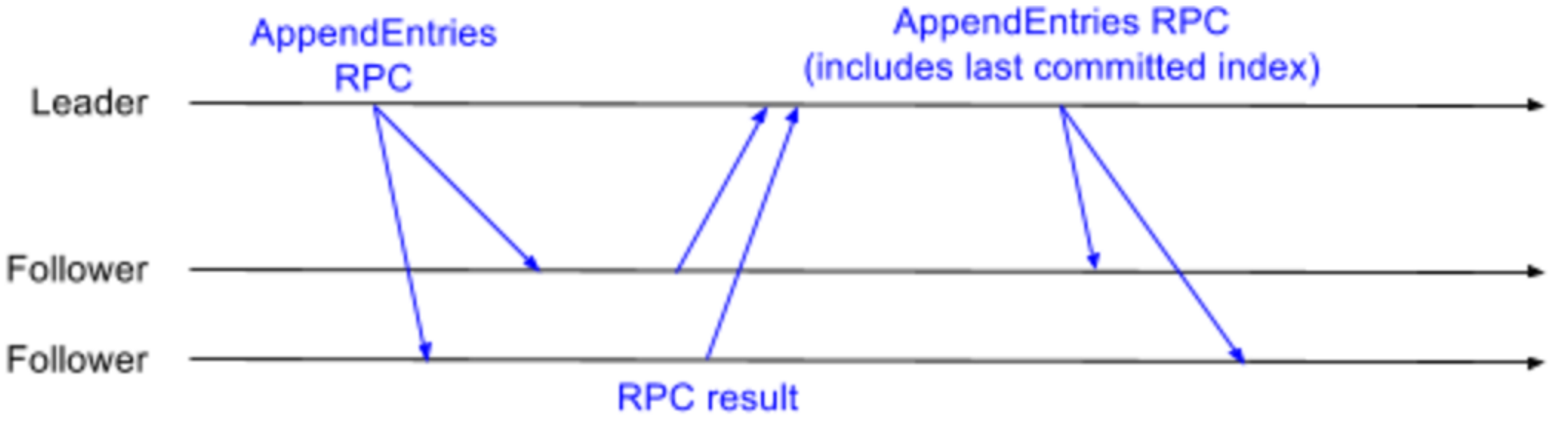


Raft

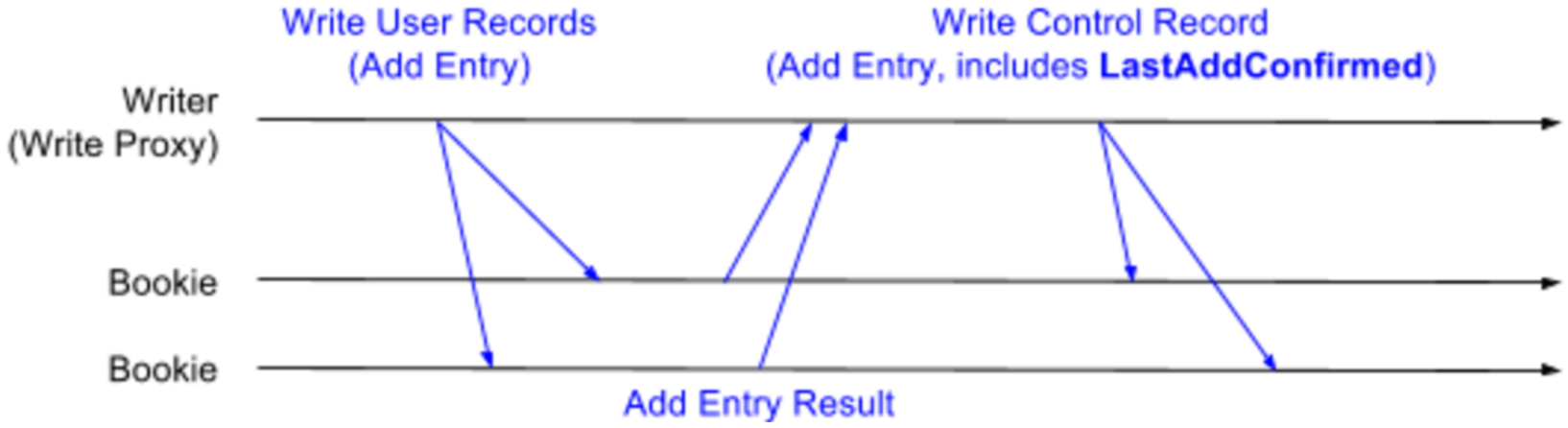
BK

资料整理
火龙果软件
荐

一致性(3)



Raft

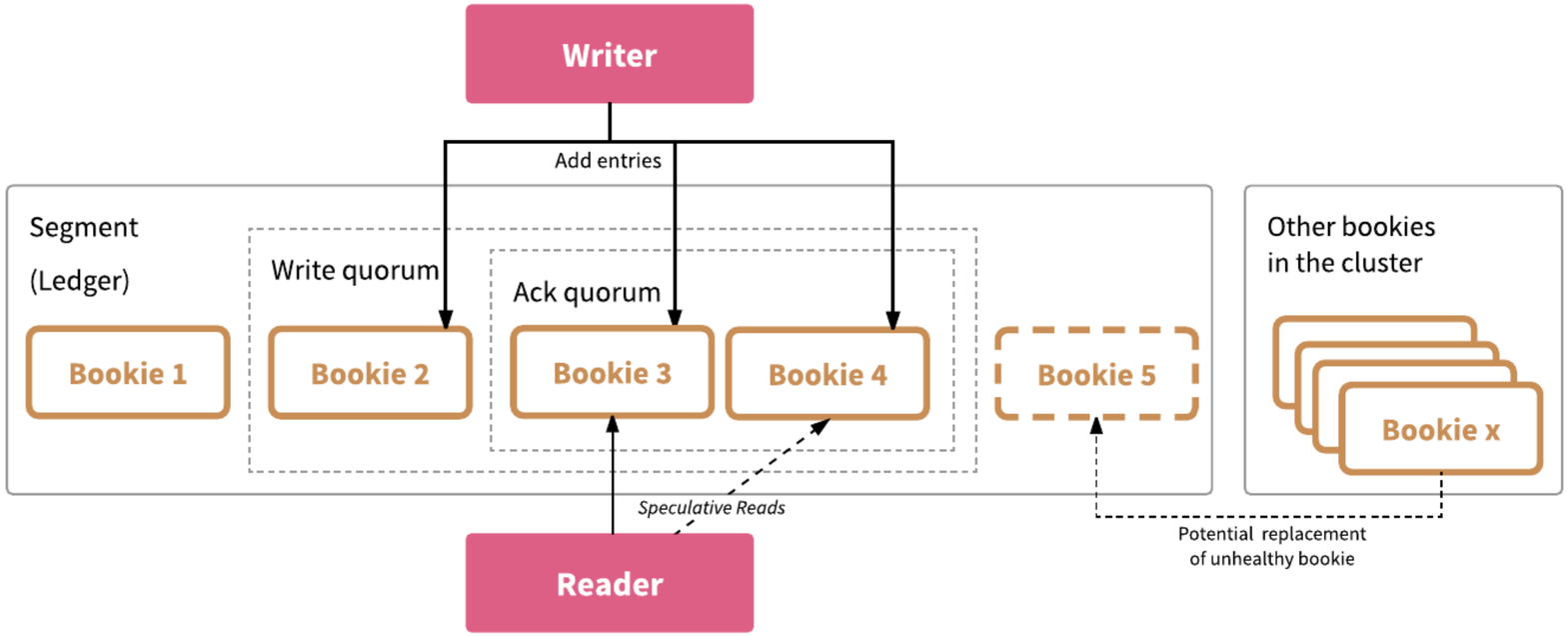


BK

资料整理
火龙果软件
荐

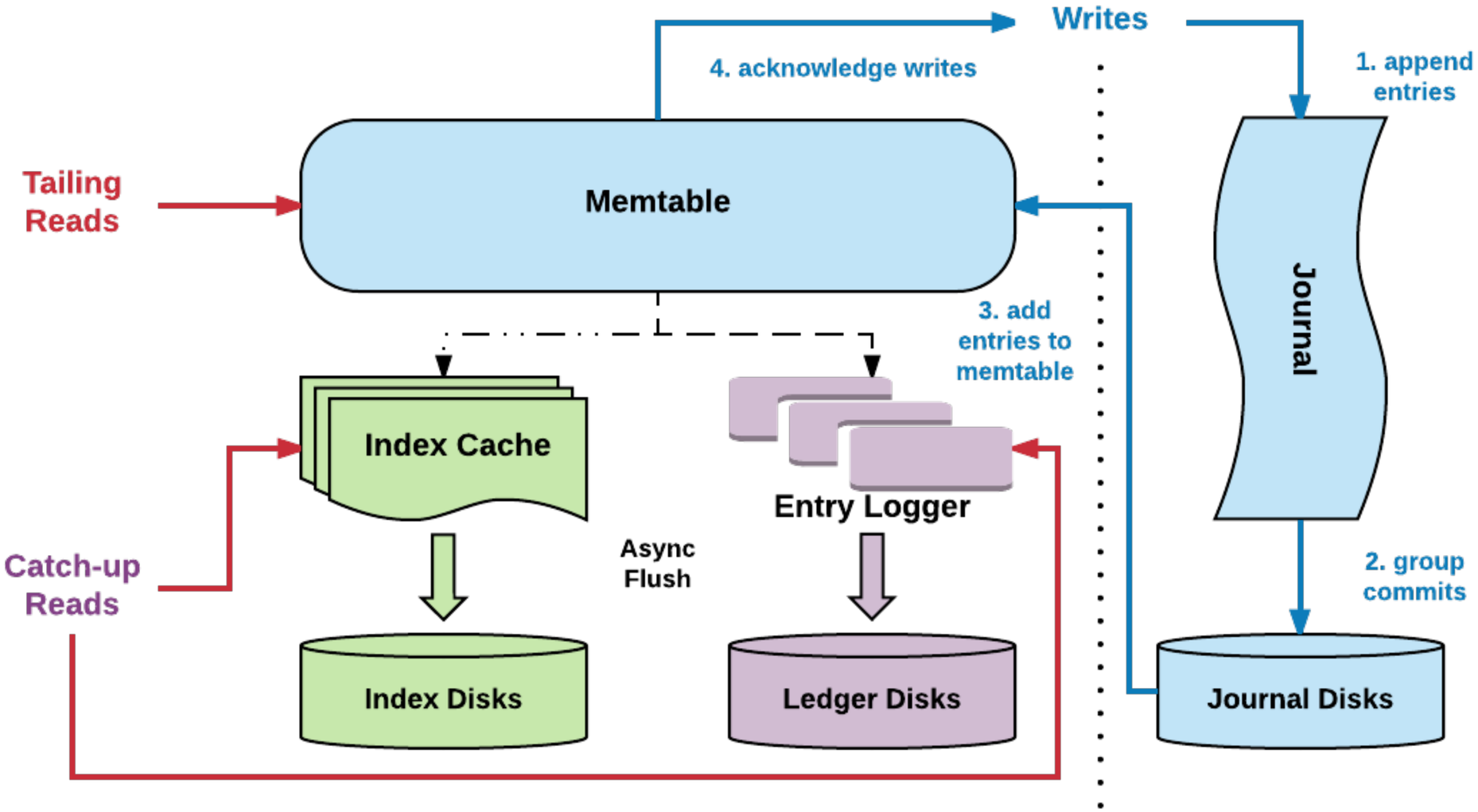
读写高可用

- 写高可用 - Ensemble Change
  - 最大化数据放置可能性
- 读高可用 - Speculative Reads
  - 没有主节点
  - 每个副本都可以提供读
  - 通过Speculative减少长尾时延



资料整理
火龙果软件
荐

存储节点的读写隔离



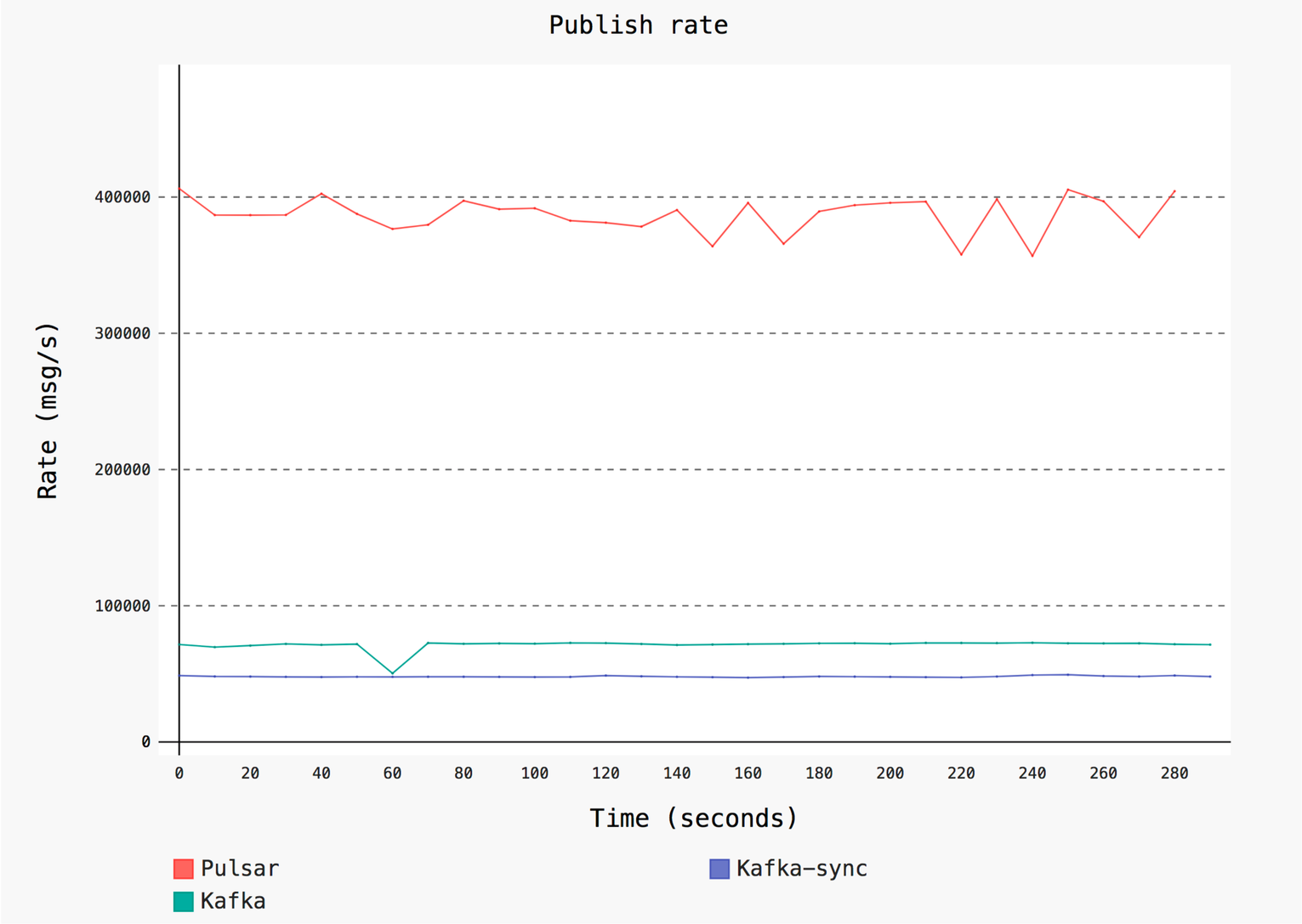
资料整理
火龙果软件
荐

Benchmark

https://github.com/openmessaging/openmessaging-benchmark

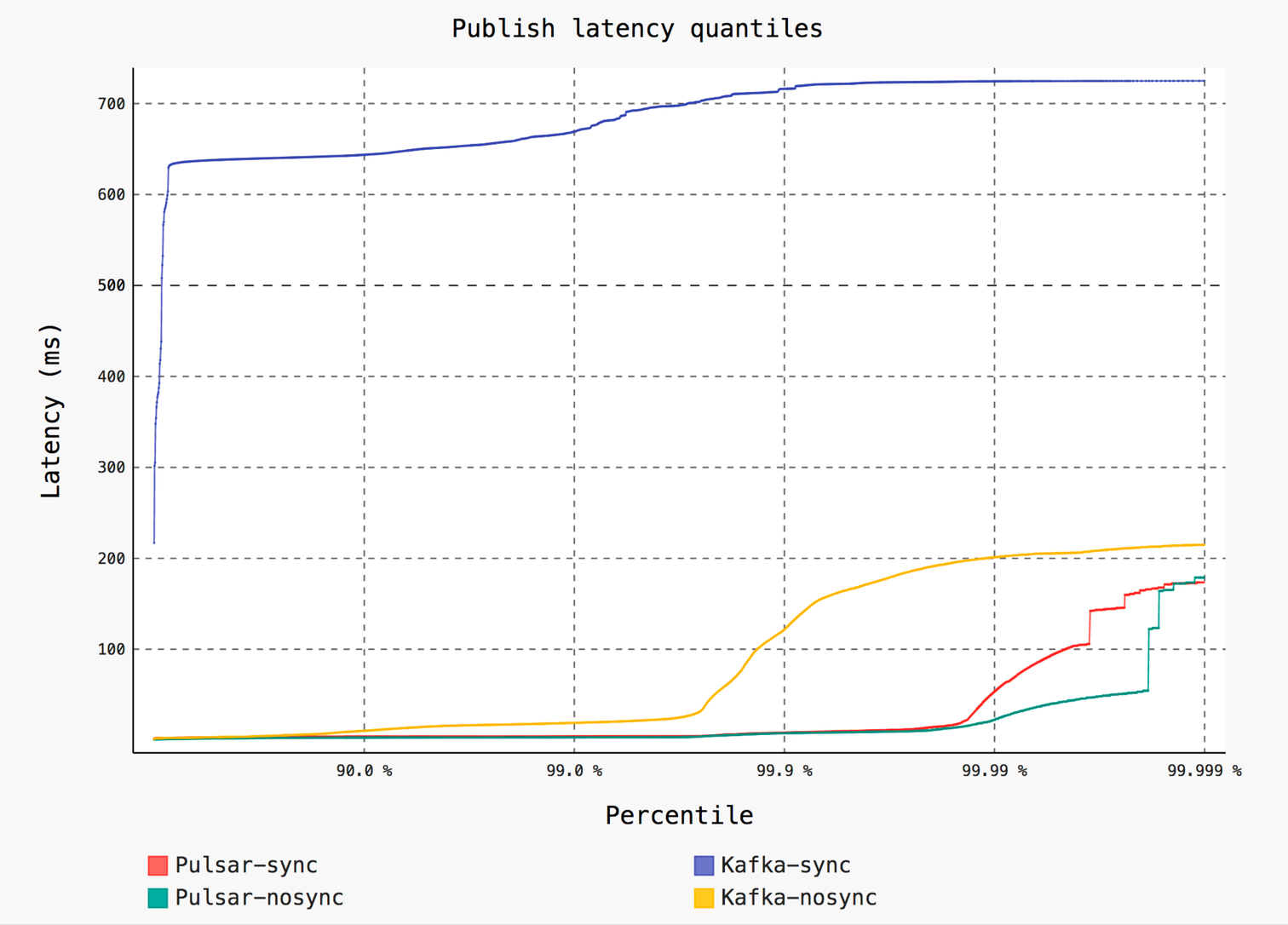
资料整理
火龙果软件
荐

Throughput



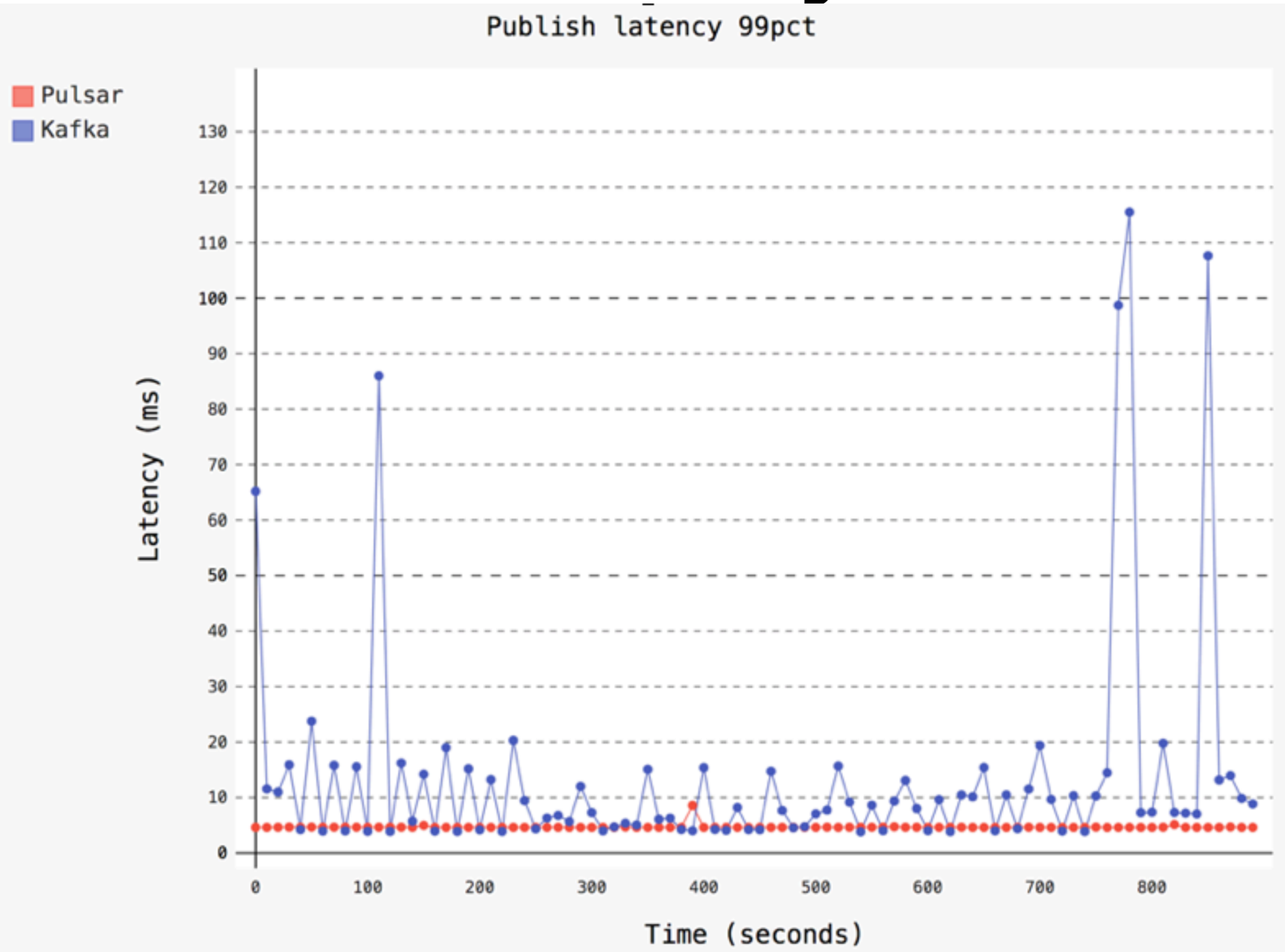
资料整理
火龙果软件
荐

Latency



资料整理
火龙果软件
荐

Latency

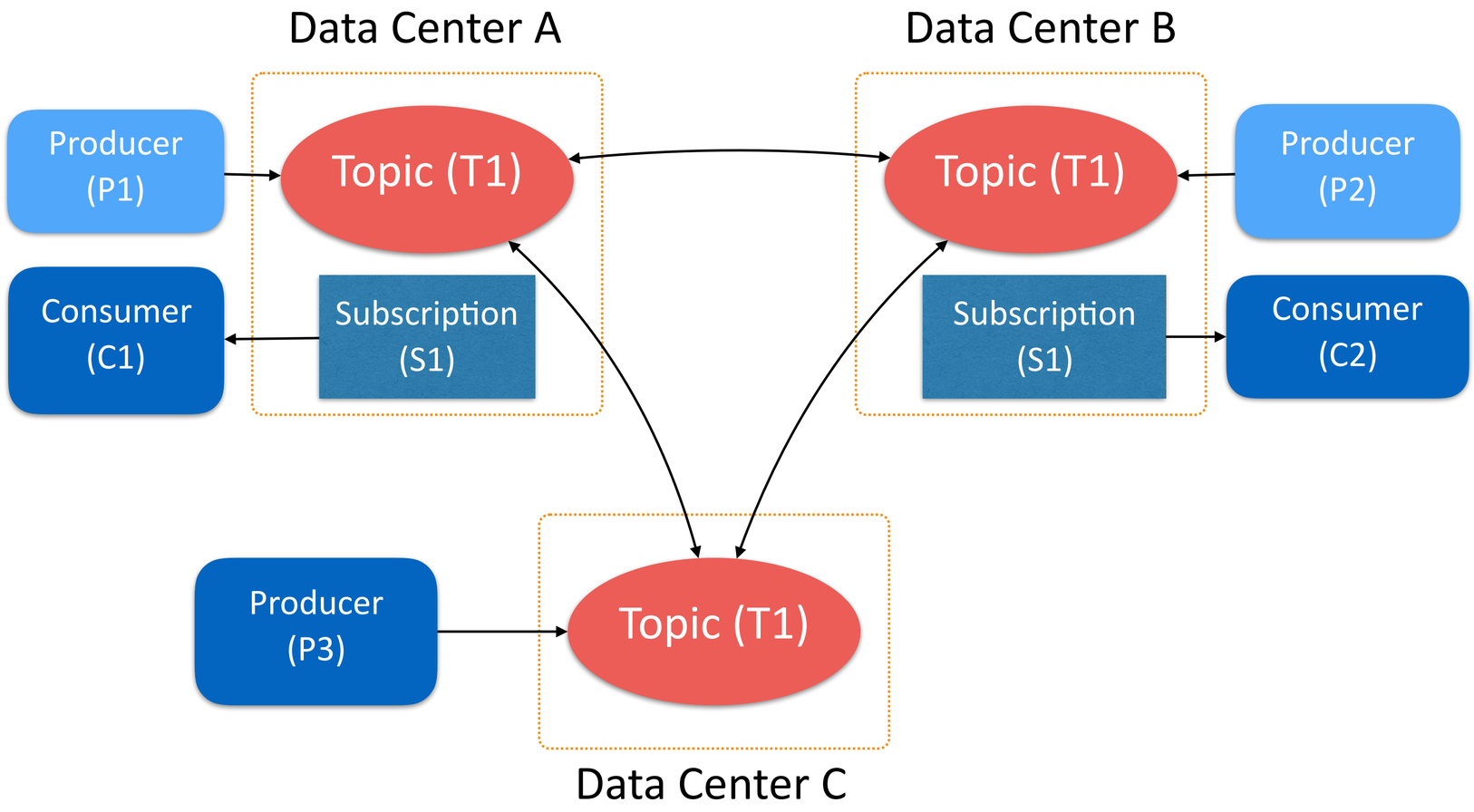


资料整理
火龙果软件
荐

谢谢

资料整理
火龙果软件
荐

多机房互备



- Scalable asynchronous replication
- Integrated in the broker message flow
- Simple configuration to add/remove regions

资料整理
火龙果软件
荐

多租户

- Authentication / Authorization / Namespaces / Admin APIs
- I/O Isolations between writes and reads
  - Provided by BookKeeper - Ensure readers draining backlog won't affect publishers
- ***Soft isolation***
  - Storage quotas – flow control – rate limiting – back pressure
- ***Hardware isolation***
  - Constrain some tenants on a subset of brokers or bookies

资料整理
火龙果软件
荐