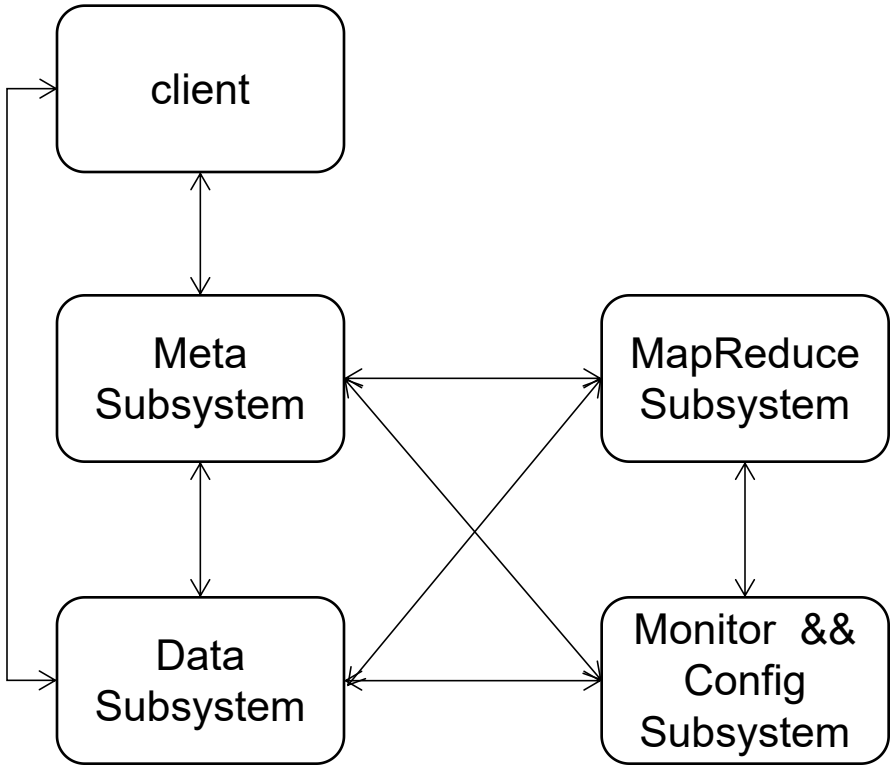
分布式存储系统

分布式文件系统 - 统一存储平台的核心

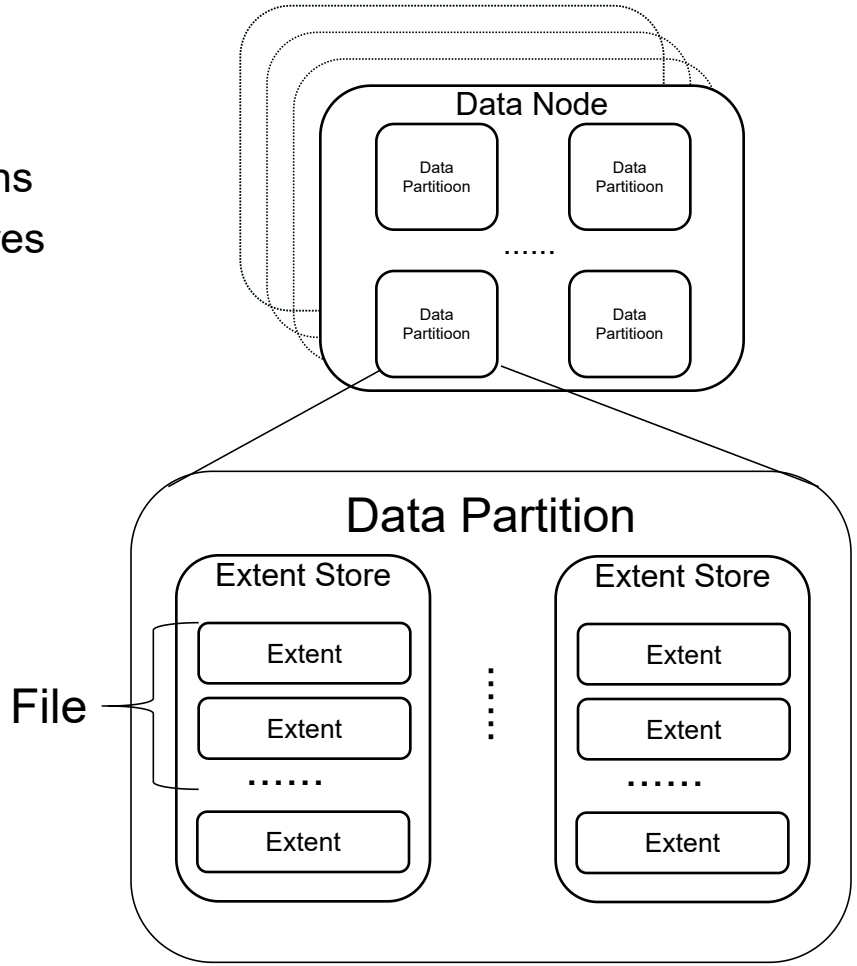
分布式文件系统 - 组件

- Client
  - 外部请求入口
- Meta Subsystem
  - 元数据管理
- Data Subsystem
  - 数据管理
- Monitor && Config Subsystem
  - 集群监控和配置
- MapReduce Subsystem
  - 集群任务调度



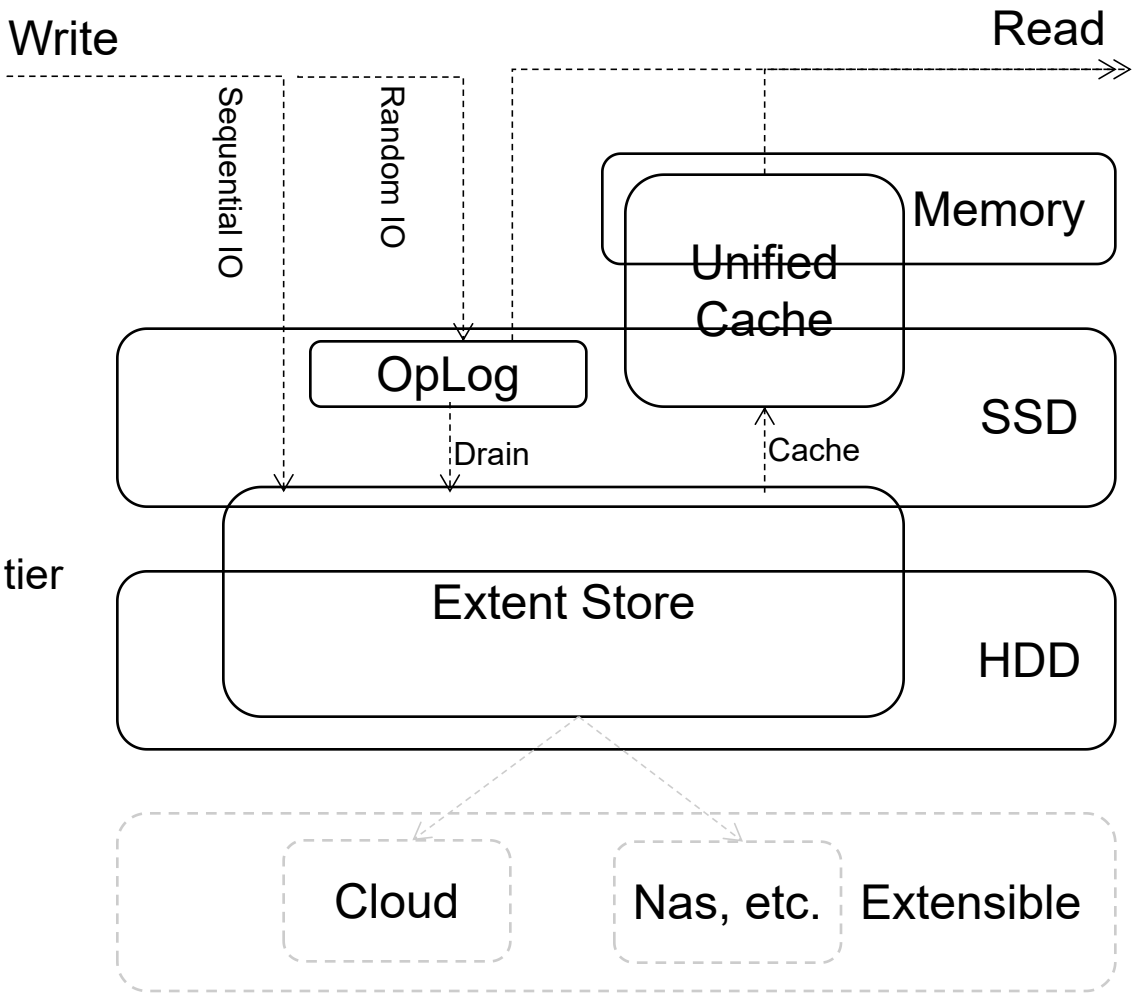
分布式文件系统 - Data Subsystem

- 逻辑结构
  - 集群包含多个data nodes
  - 每个data nodes包含多个data partitions
  - 每个data partition包含多个extent stores
  - 每个extent store包含多个extents
- 文件
  - 每个文件由一个或者多个extents组成
  - 多个小文件聚合共用同一个extent
- Extent
  - 每个extent维护多个副本

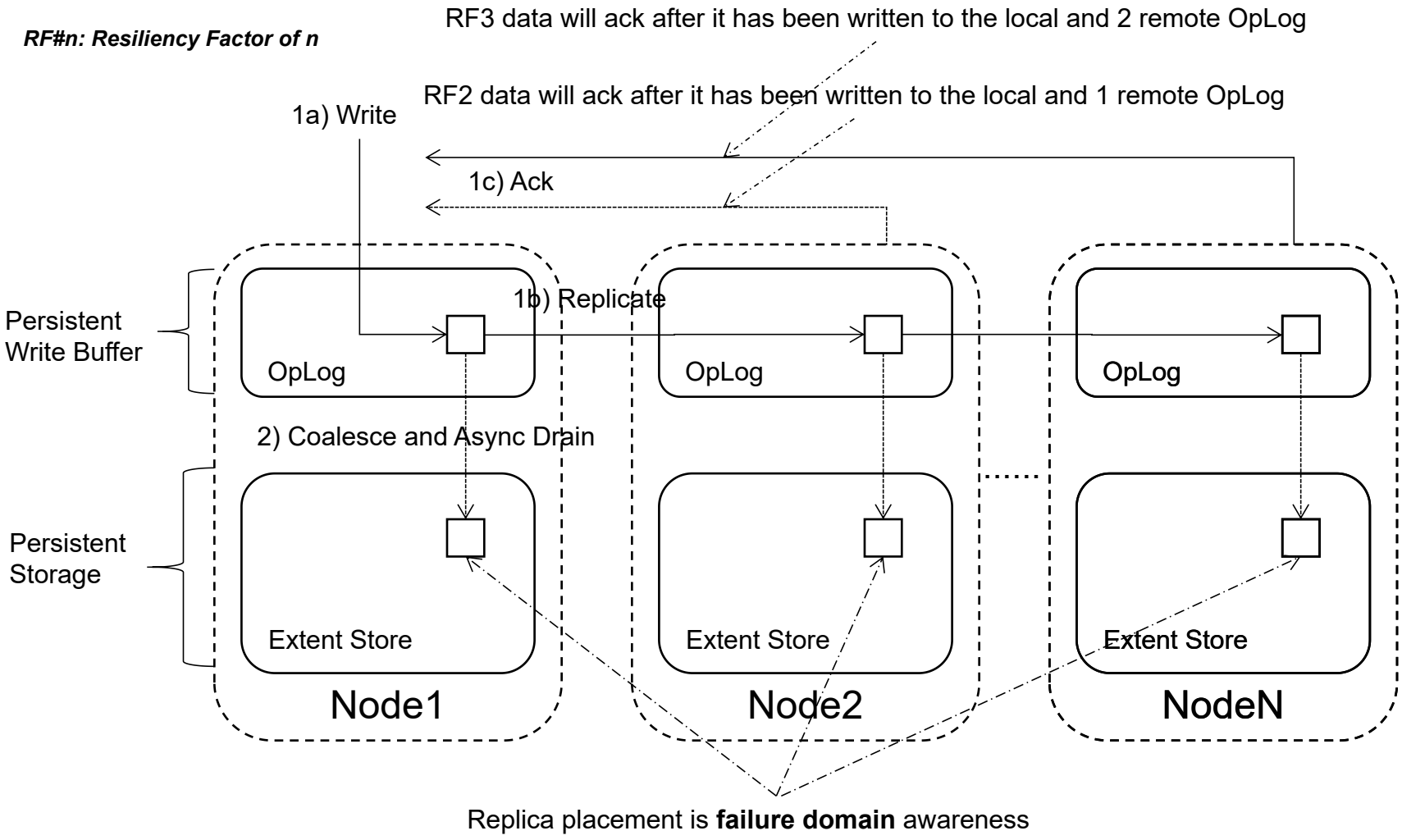


Data Subsystem - IO Path and Cache

- OpLog
  - 持久化的write buffer
  - journal structure
  - 用于random IO
  - SSD only
- Extent Store
  - 持久化的数据存储
  - SSD/HDD/extensible tier
- Unified Cache
  - SSD/Memory

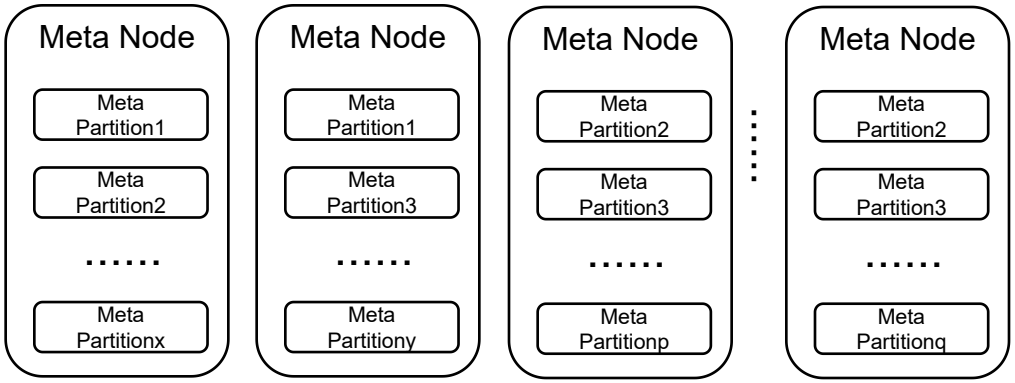


Data Subsystem - Data Protection

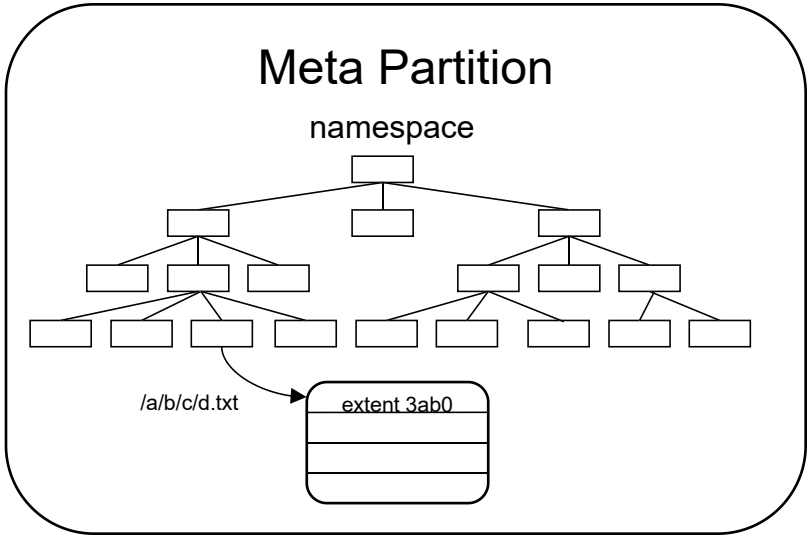


分布式文件系统 - Meta Subsystem

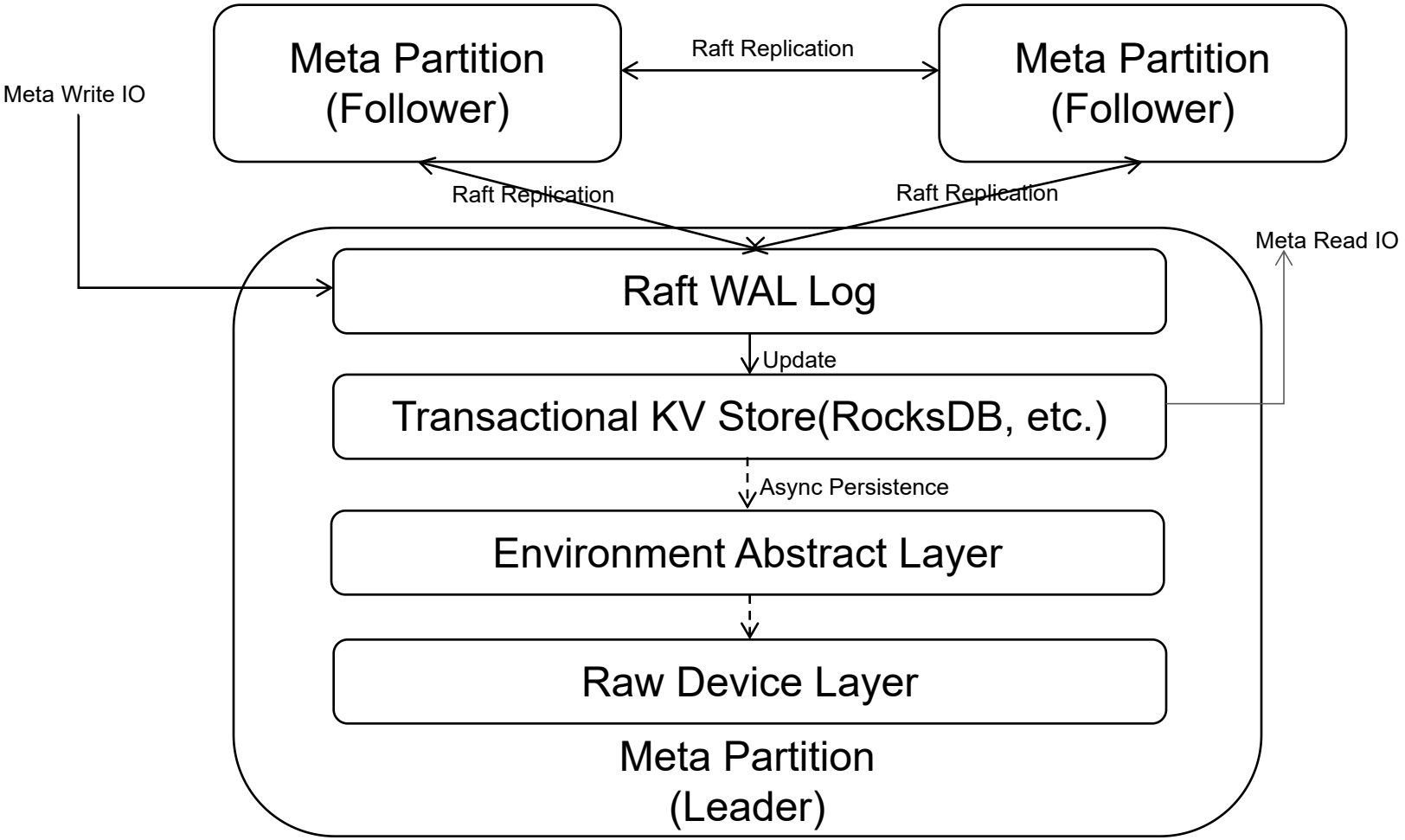
- 逻辑结构
  - 集群包含多个meta nodes
  - 每个meta nodes包含多个meta partitions



- Meta Partition
  - 每个meta partition维护局部名字空间，所有meta partitions形成全局名字空间
  - 每个meta partition维护多个副本



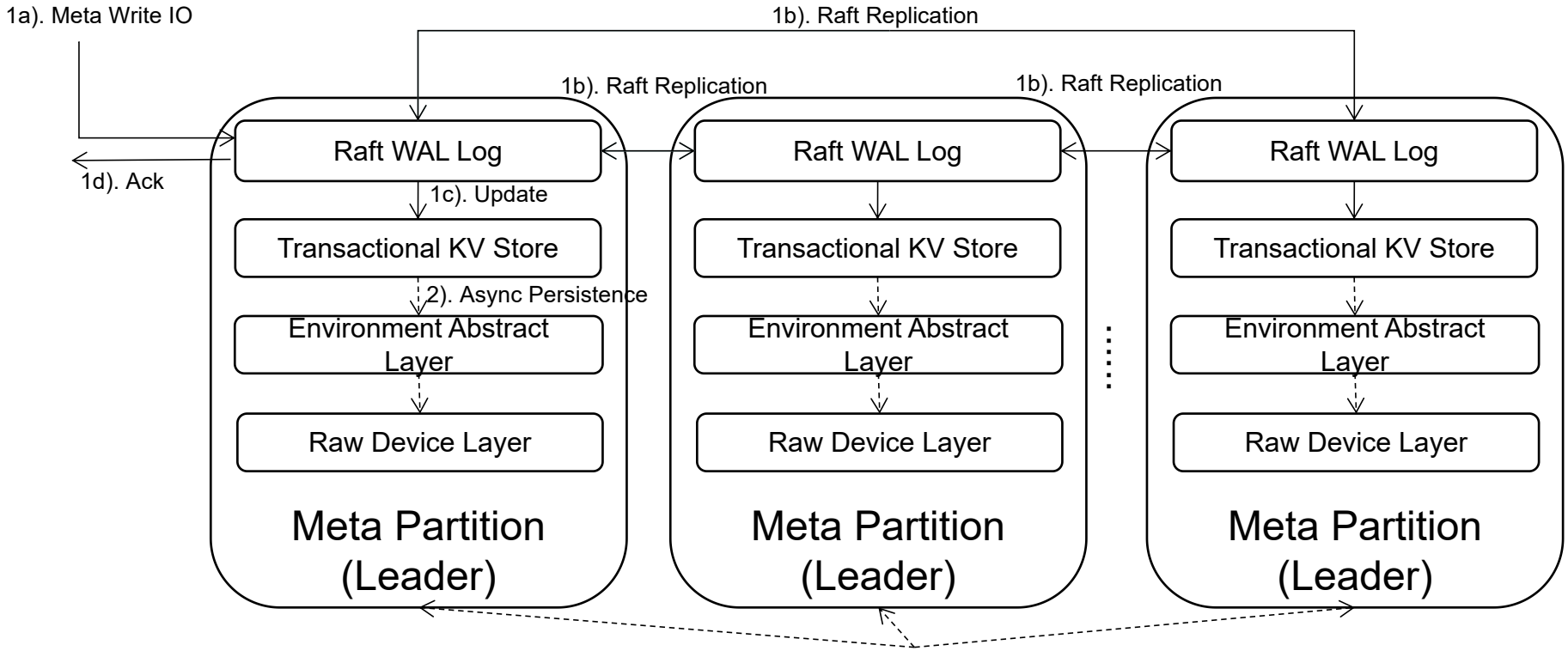
Meta Subsystem - IO Path



Meta Subsystem - Meta Protection

- Strong Consistency
  - Raft Consensus Protocal

- Transactional ACID
  - Leverage Transactional KV Store's ACID guarantee



Partition Replica placement is **failure domain** awareness