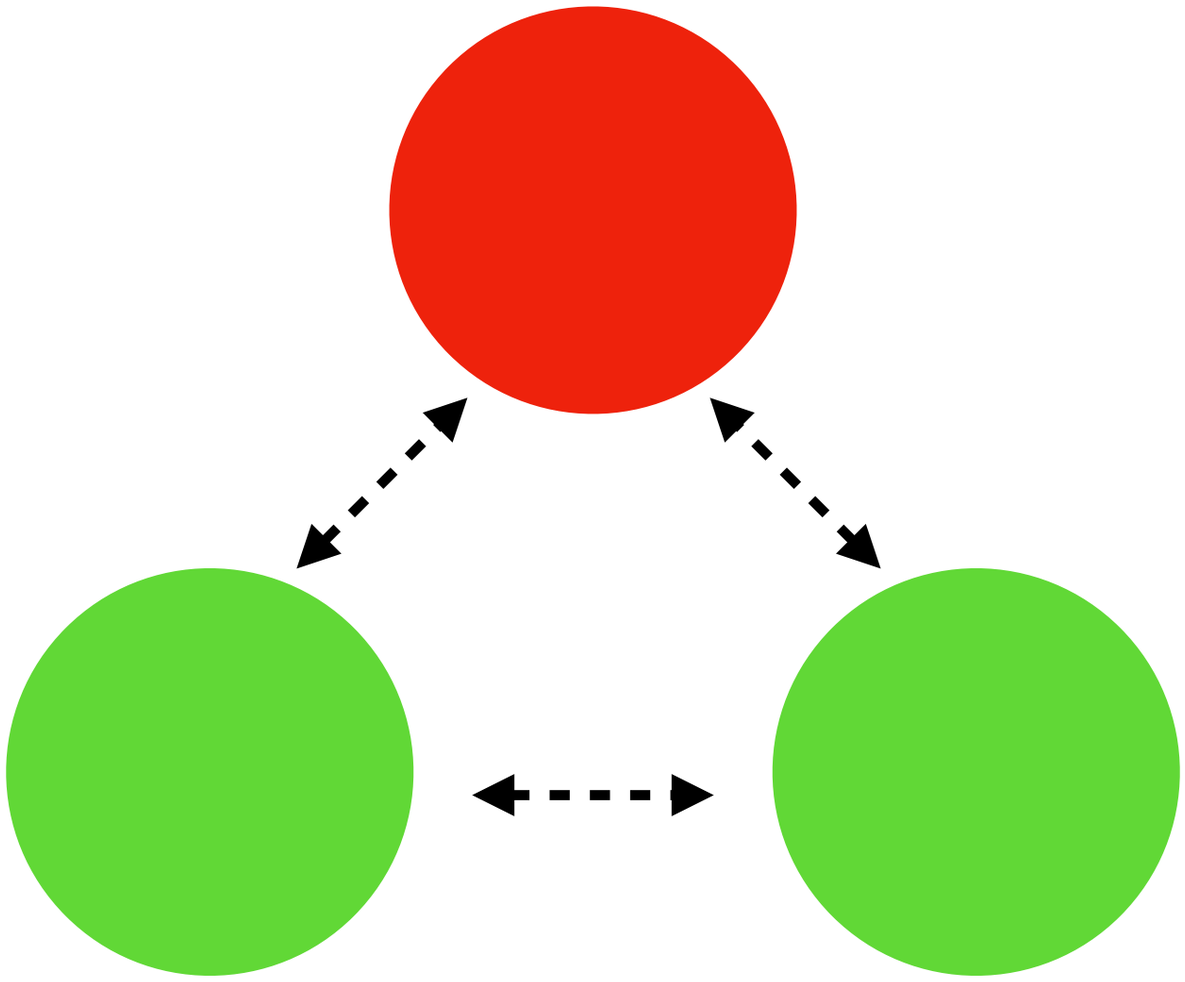


图解Raft协议

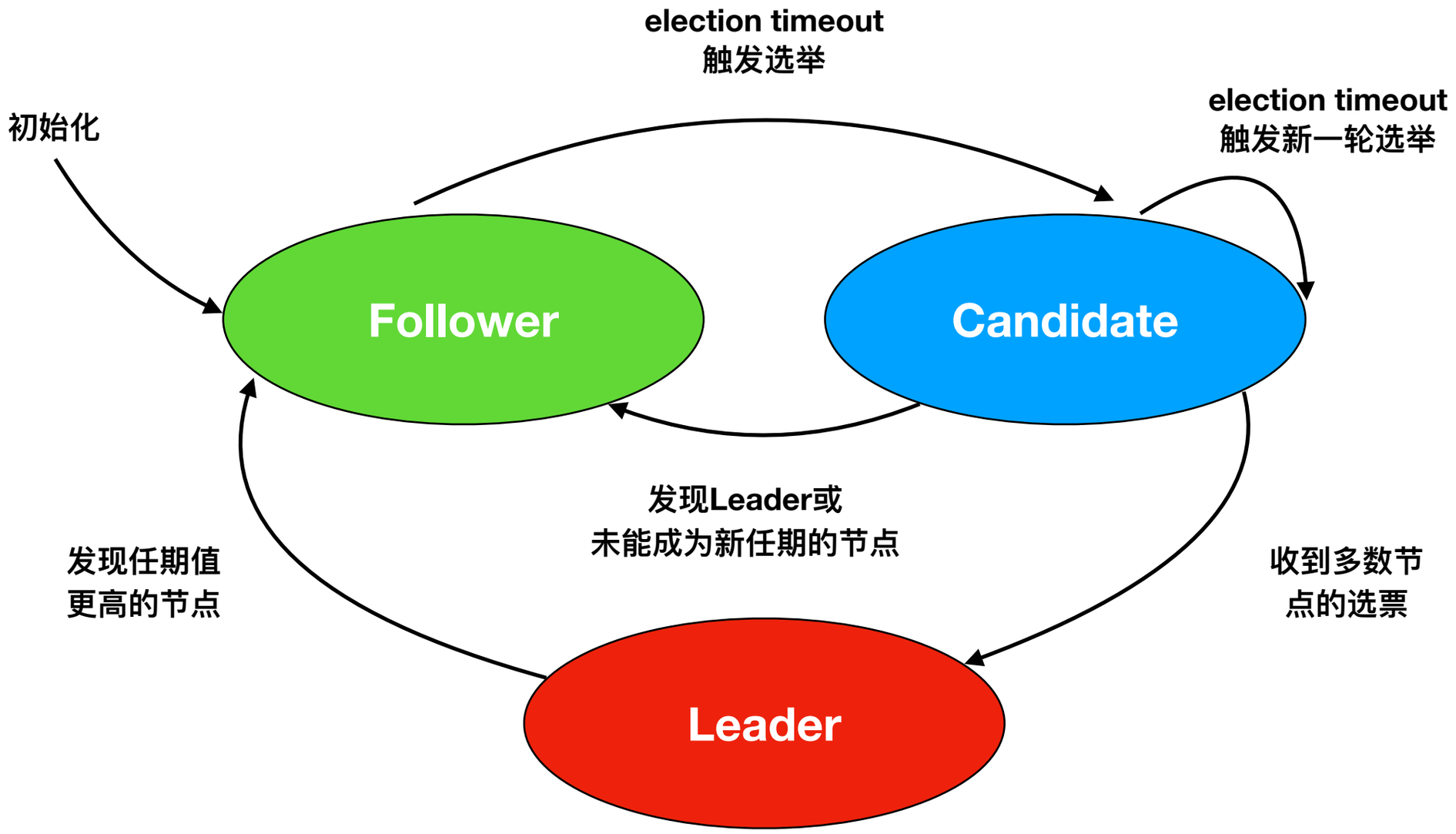
Taylor

2019-05-28

图解Raft协议

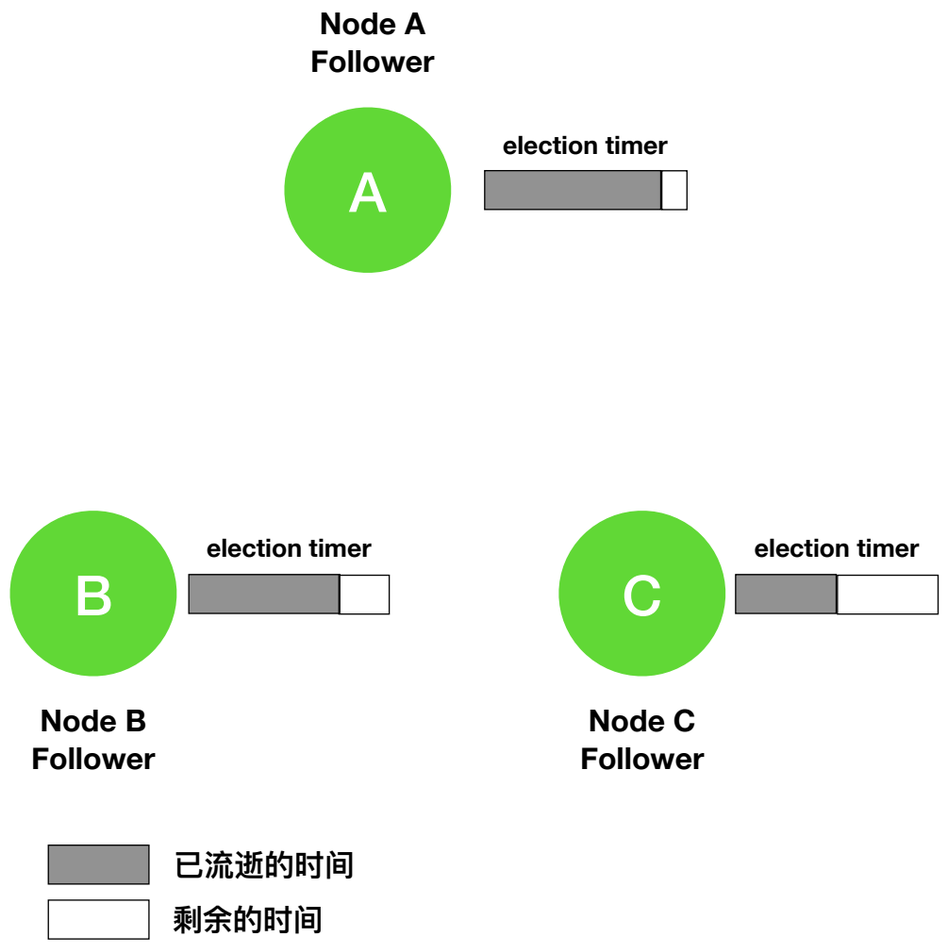
- Leader 选举
- 日志复制
- 网络分区场景
- 日志压缩与快照
- 其它技术点

Leader 选举



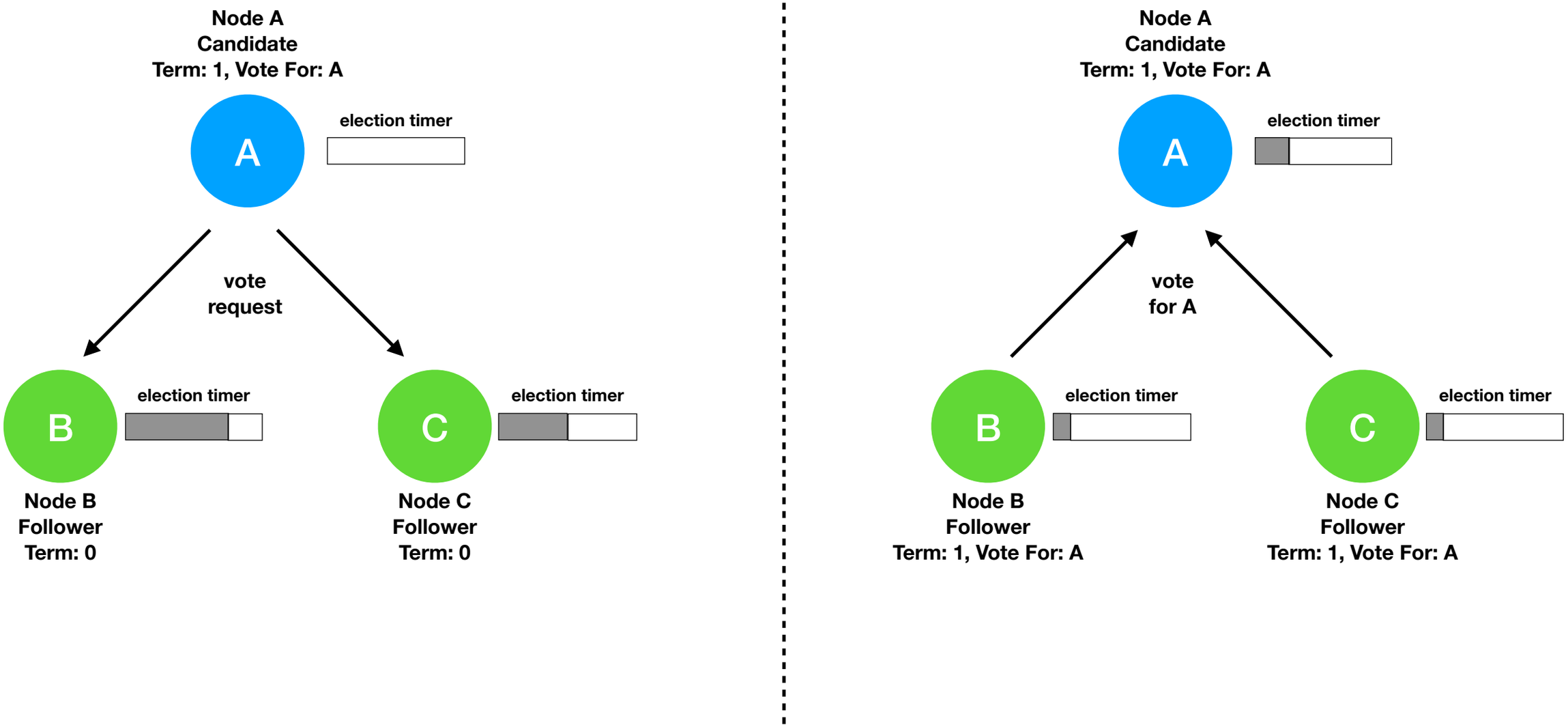
状态转换图

Leader 选举



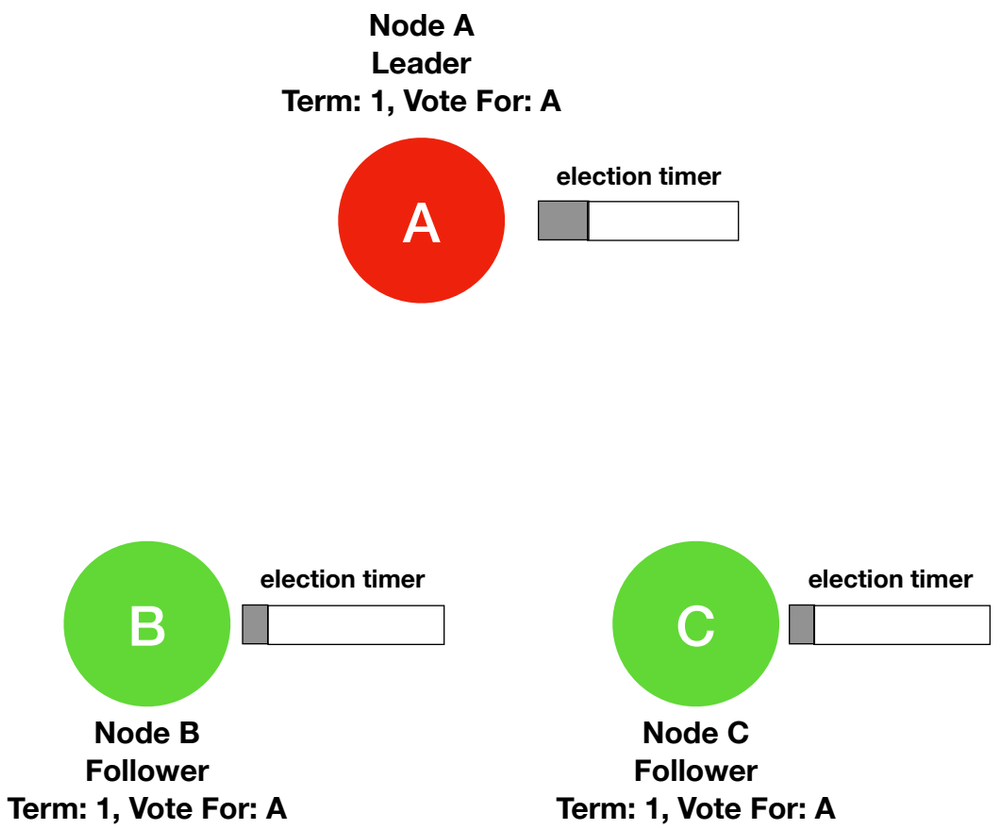
Initial State

Leader 选举



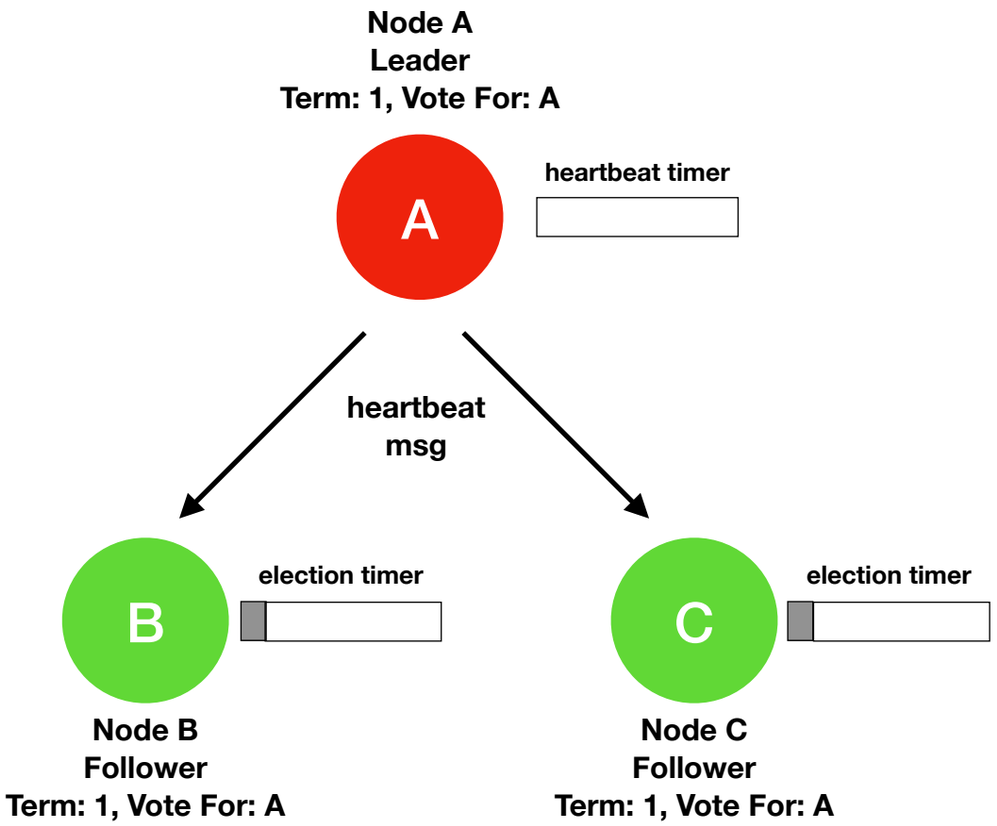
Normal Election

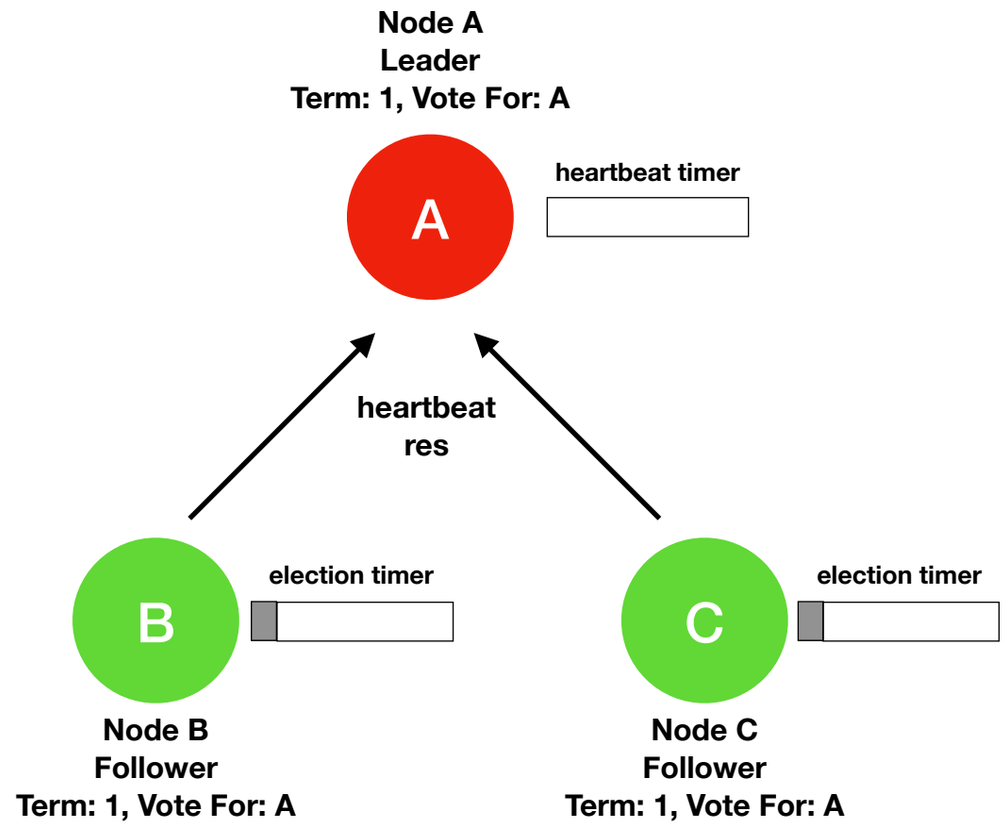
Leader 选举



Normal Election

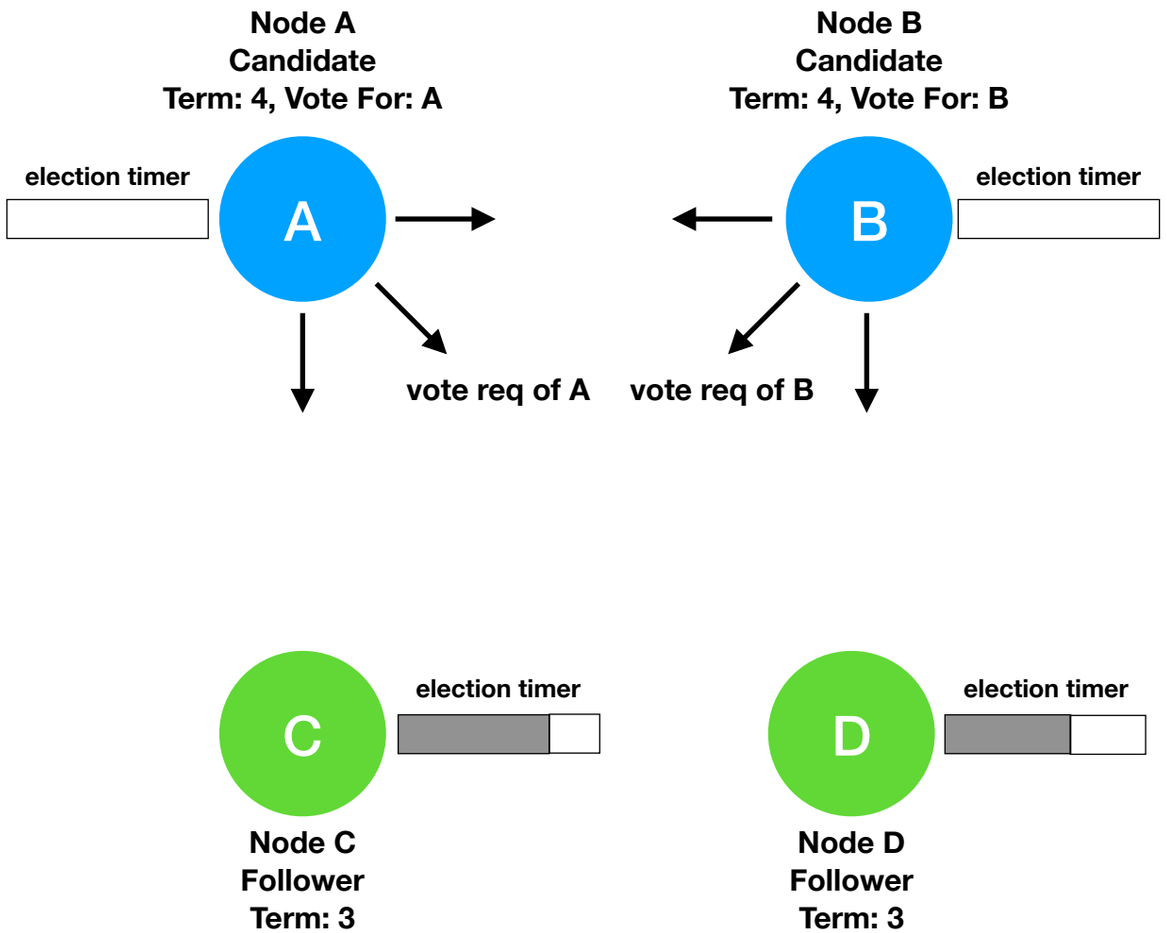
Leader 选举

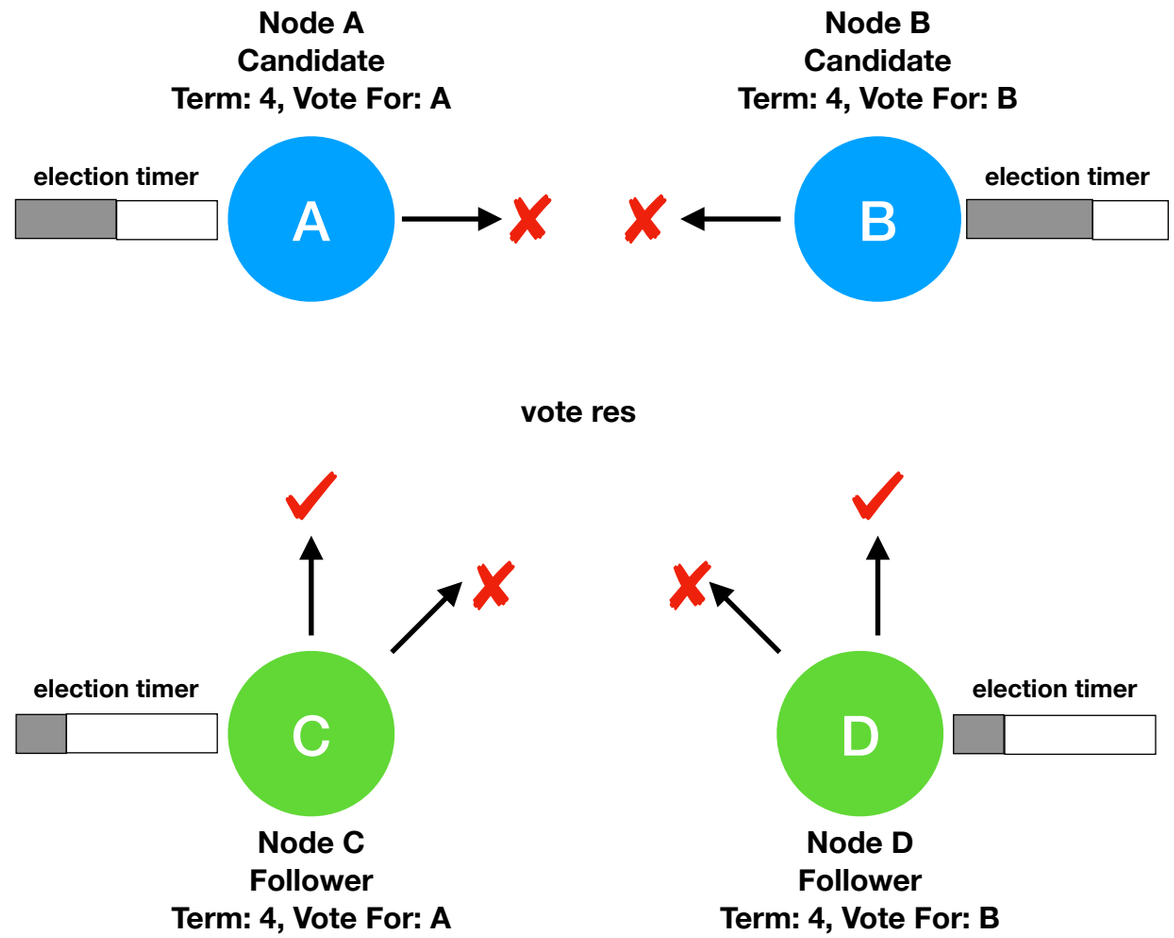




Heartbeat

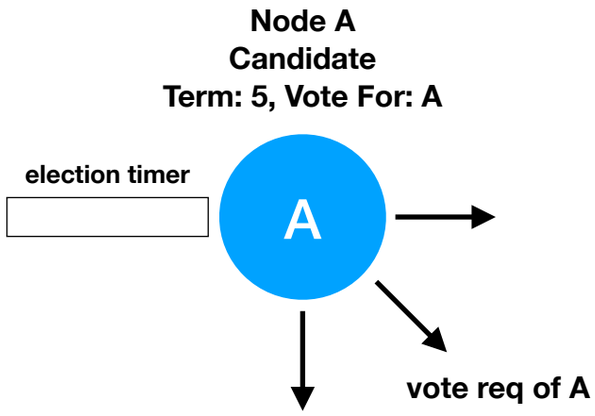
Leader 选举

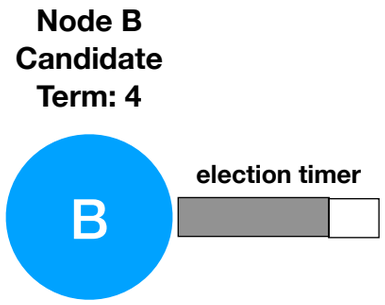


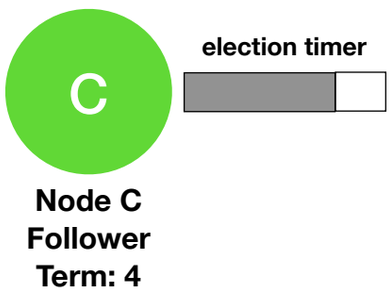


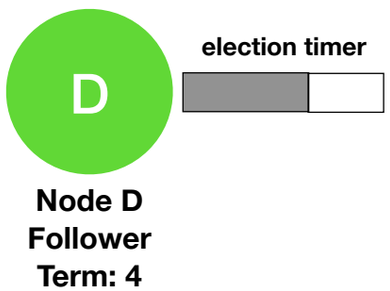
Election Conflict

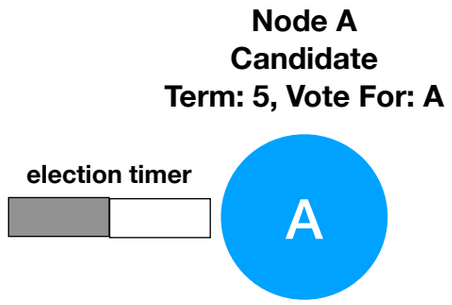
Leader 选举

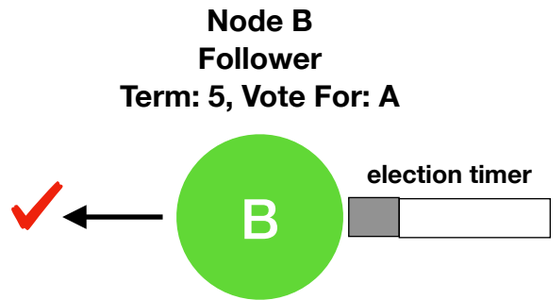


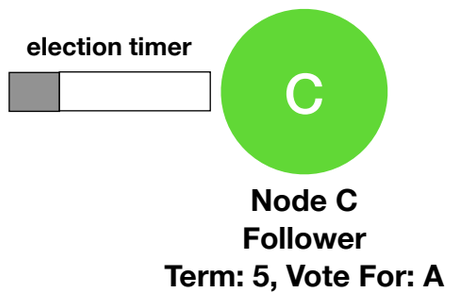


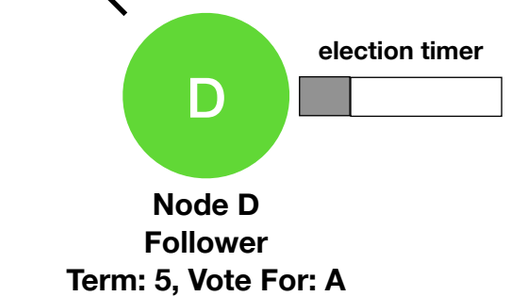






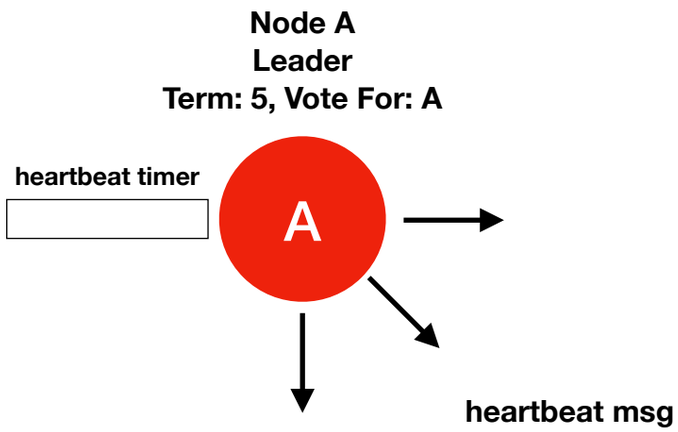


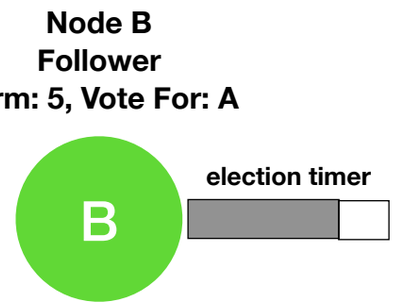


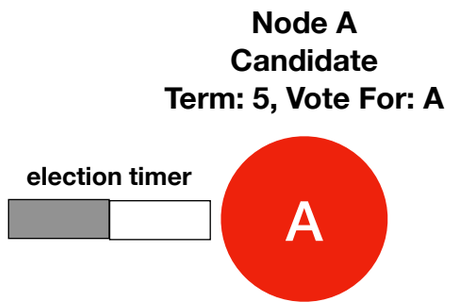


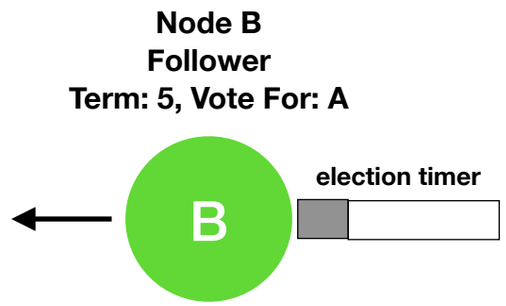
Election Conflict

Leader 选举

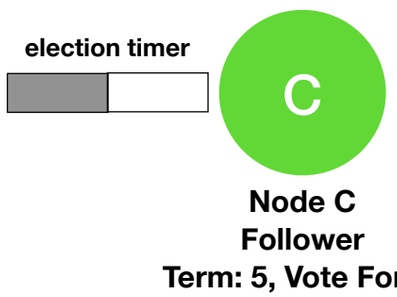


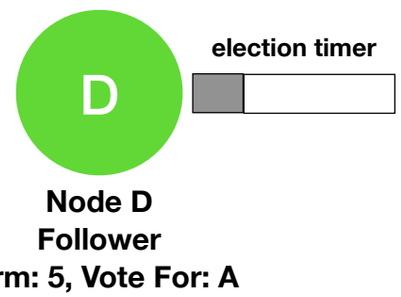


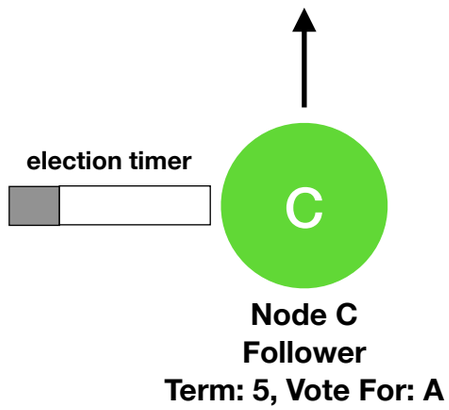


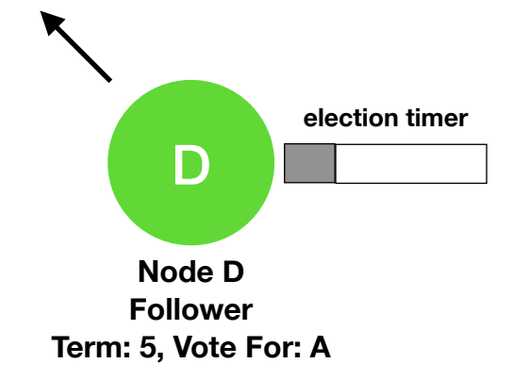


heartbeat res



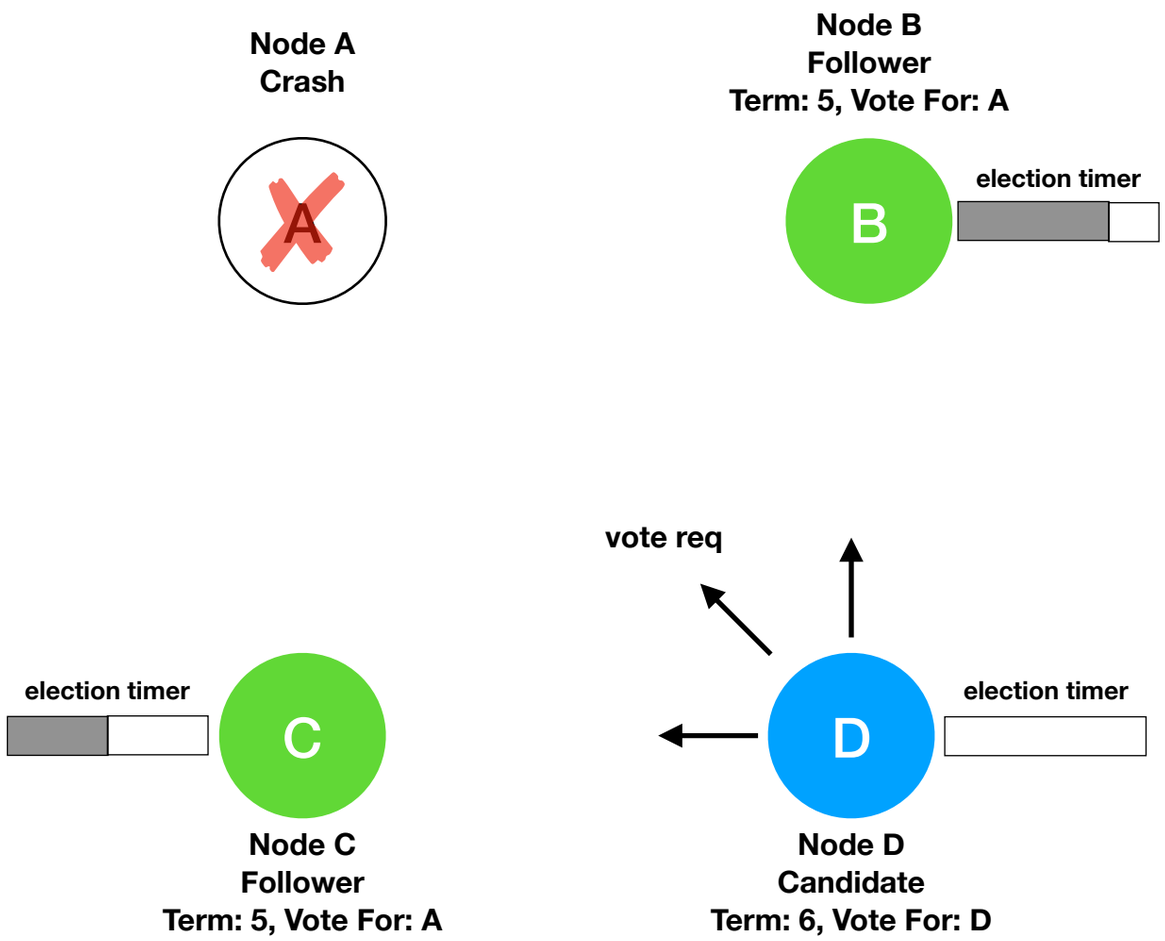


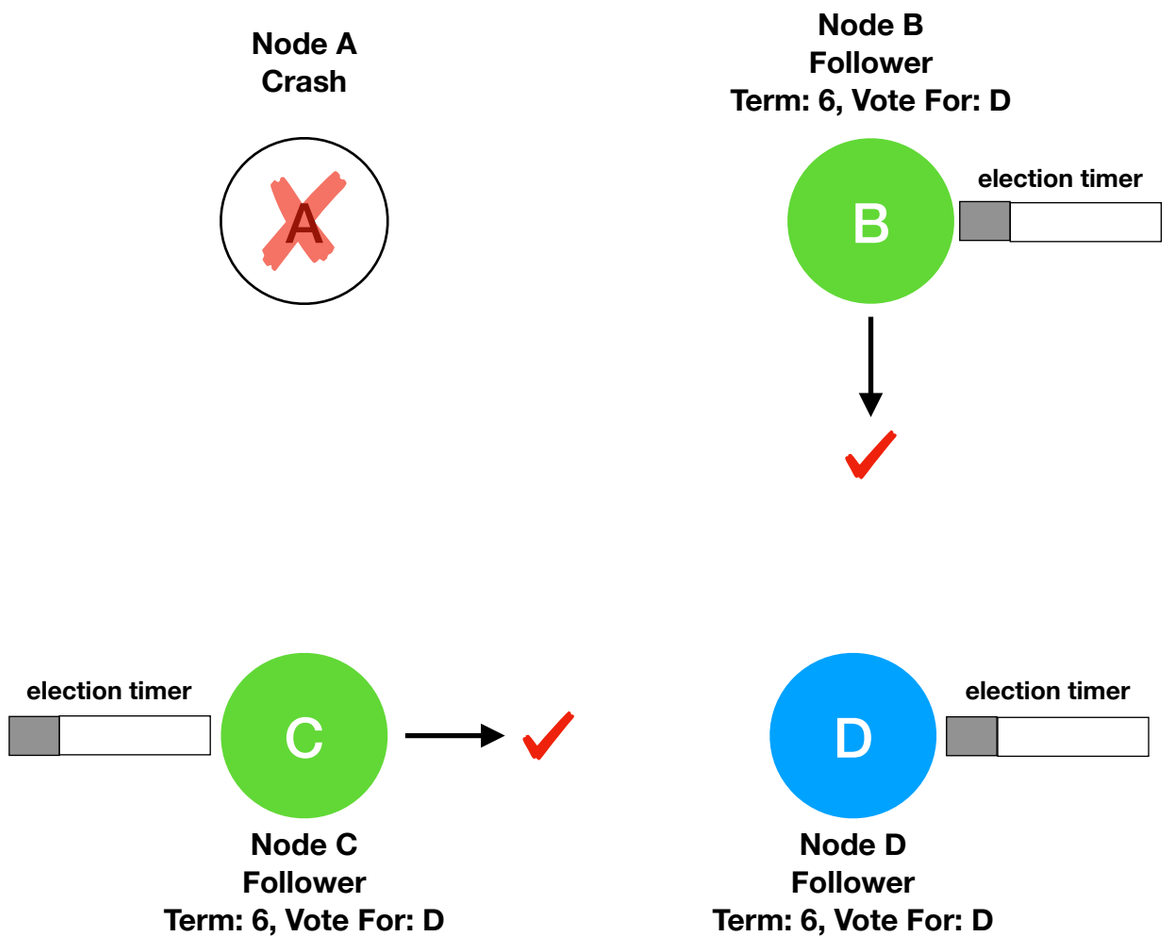




Heartbeat

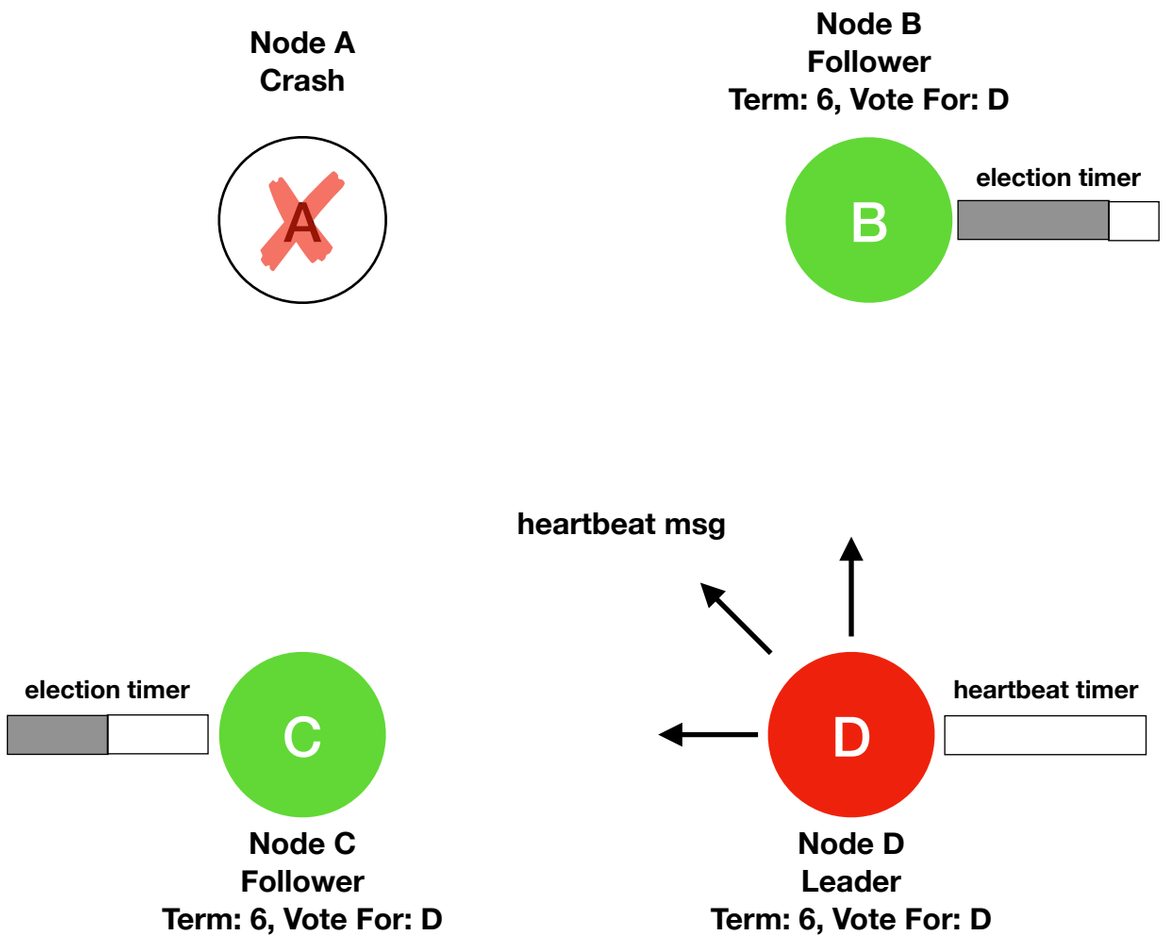
Leader 选举

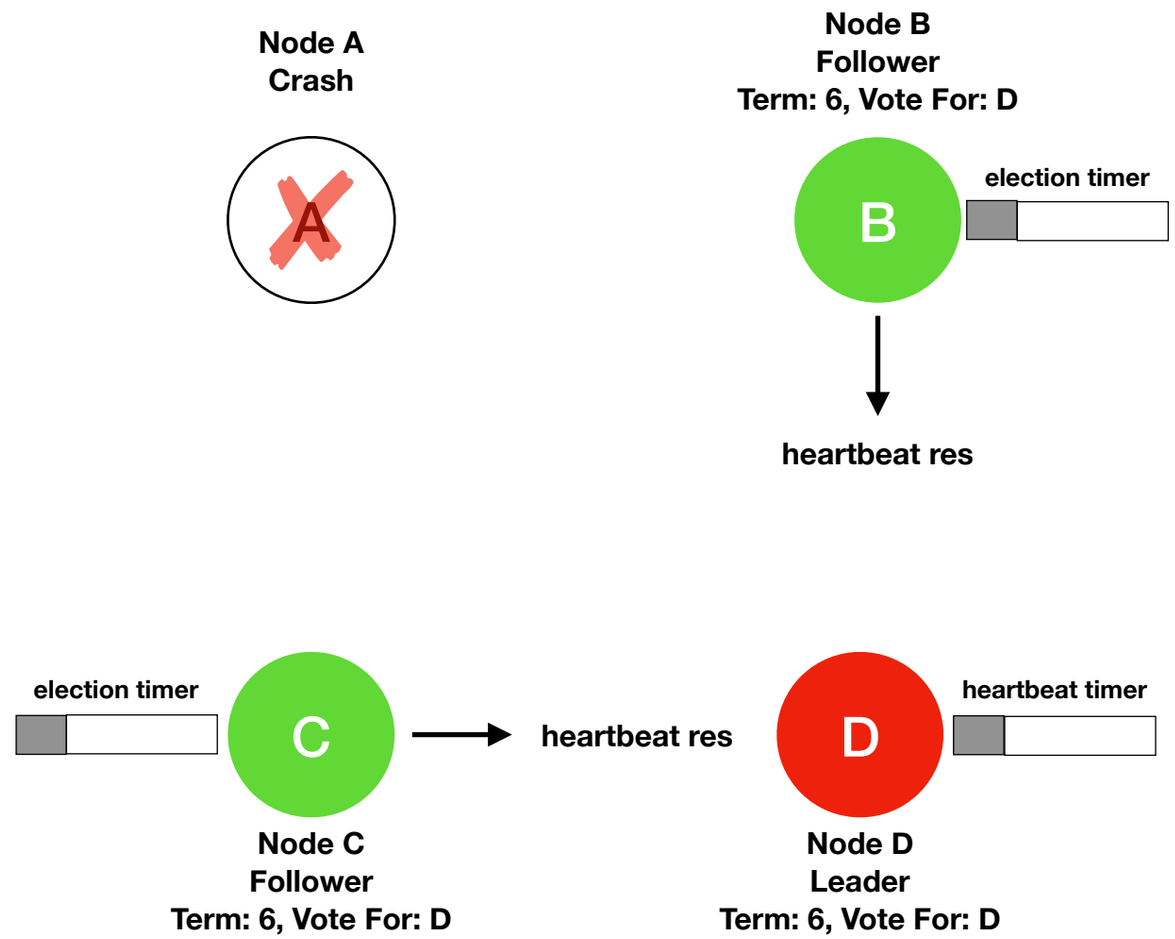




Election After Leader Crash

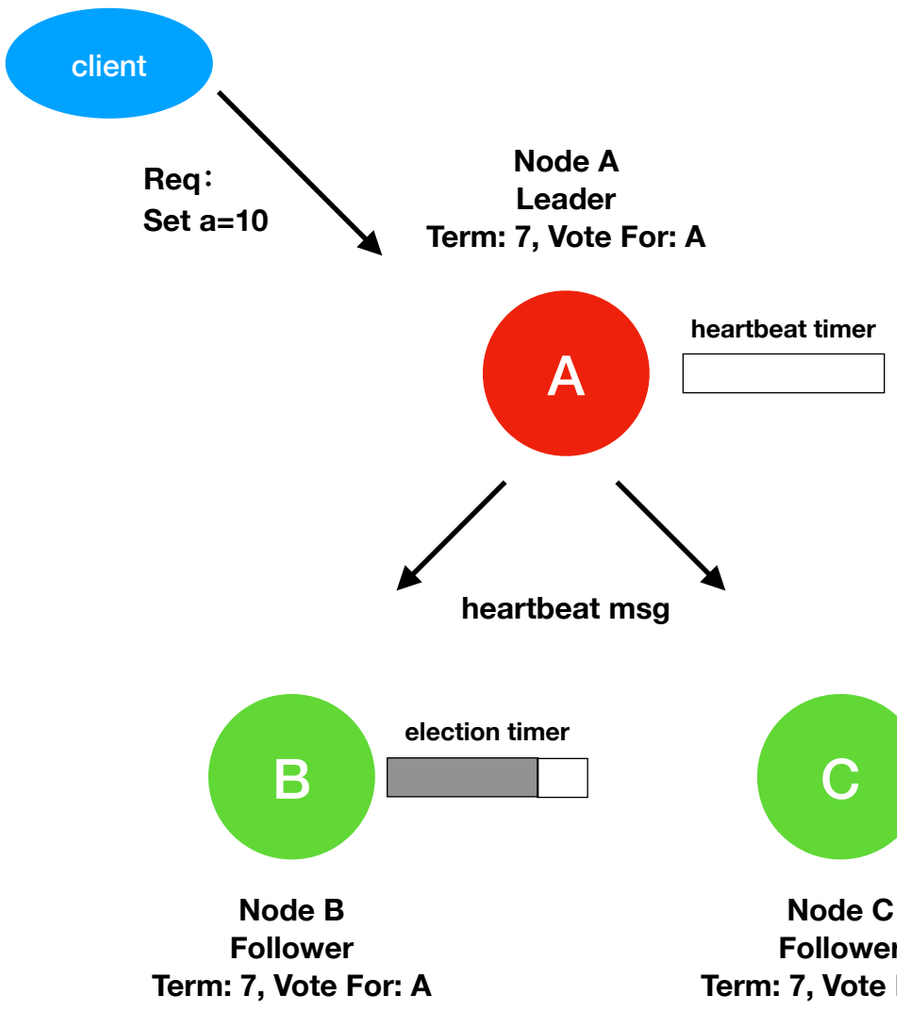
Leader 选举

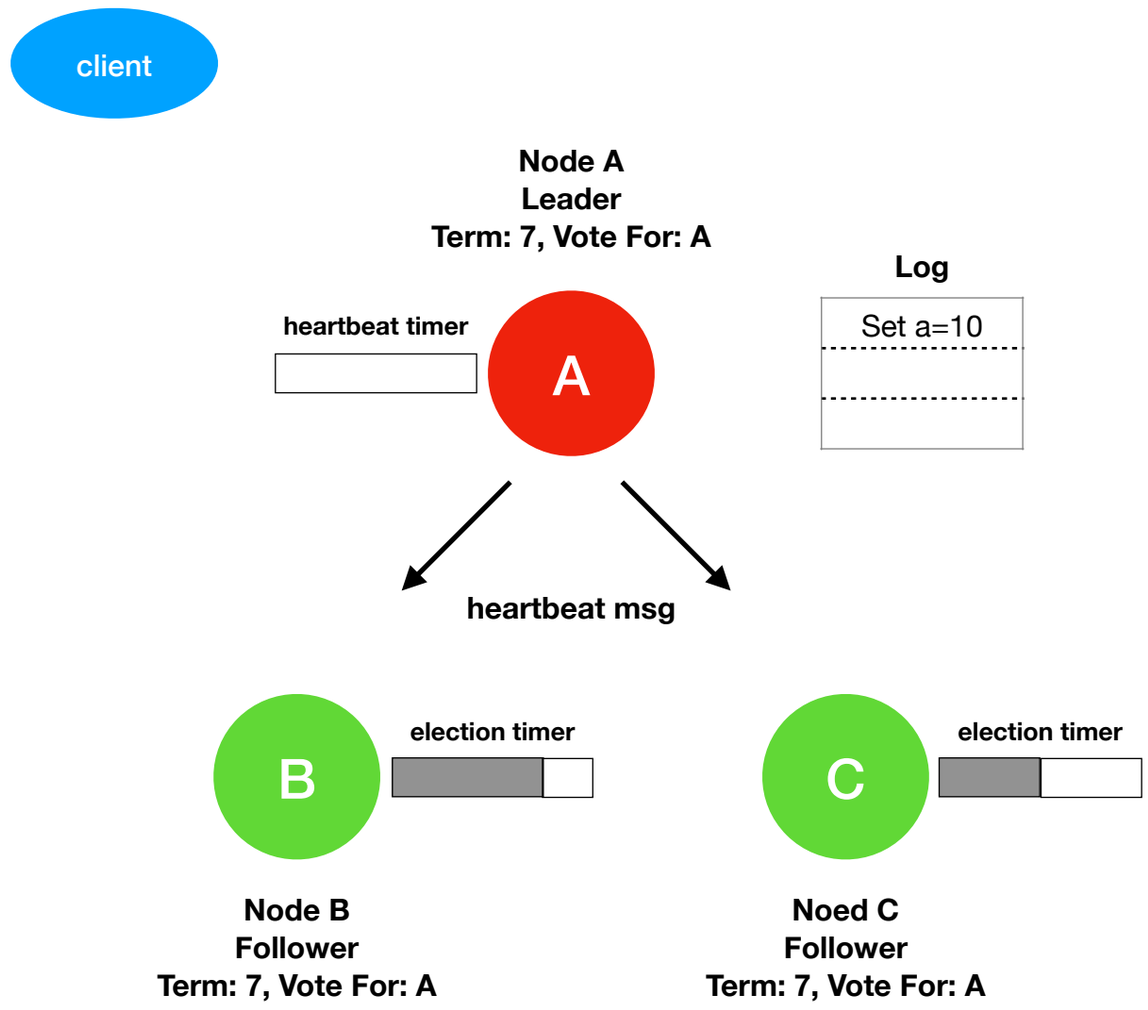




Heartbeat

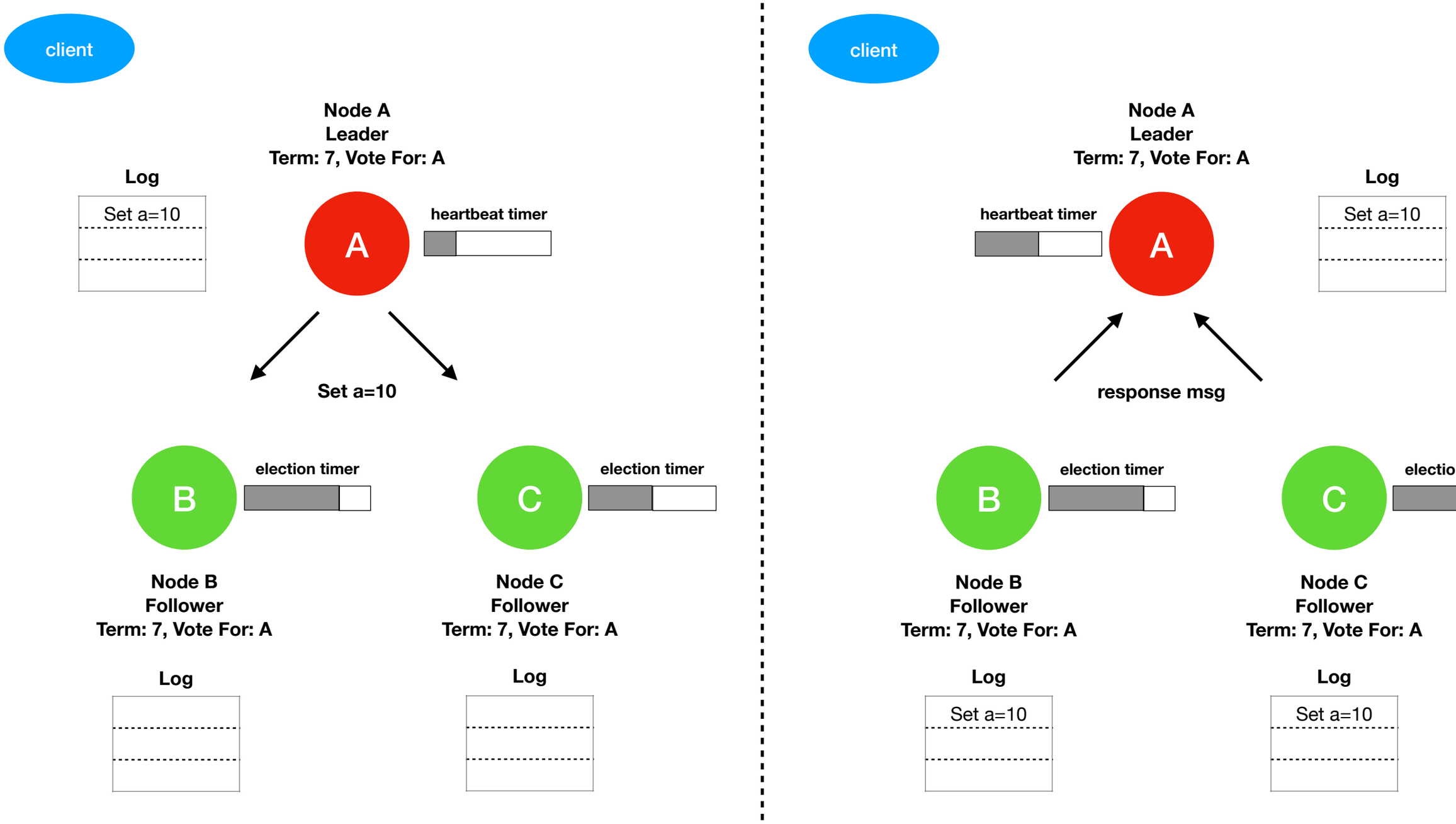
日志复制





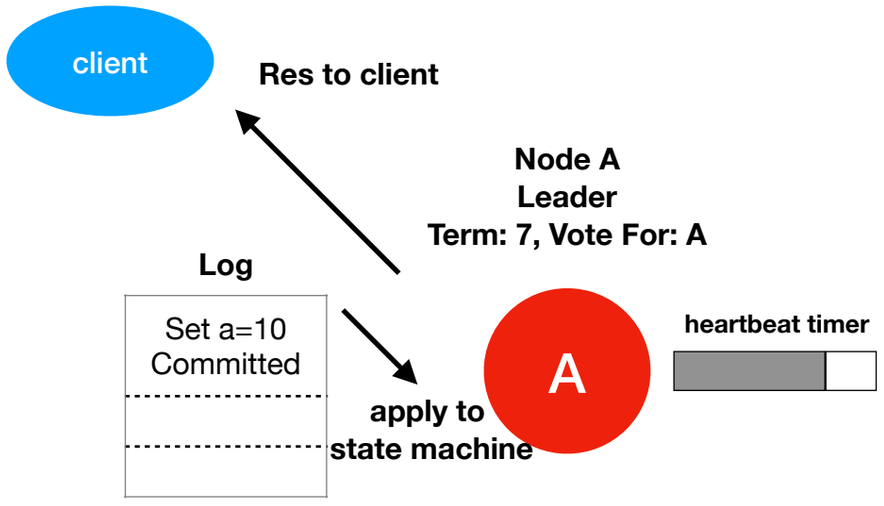
Log Replication Process

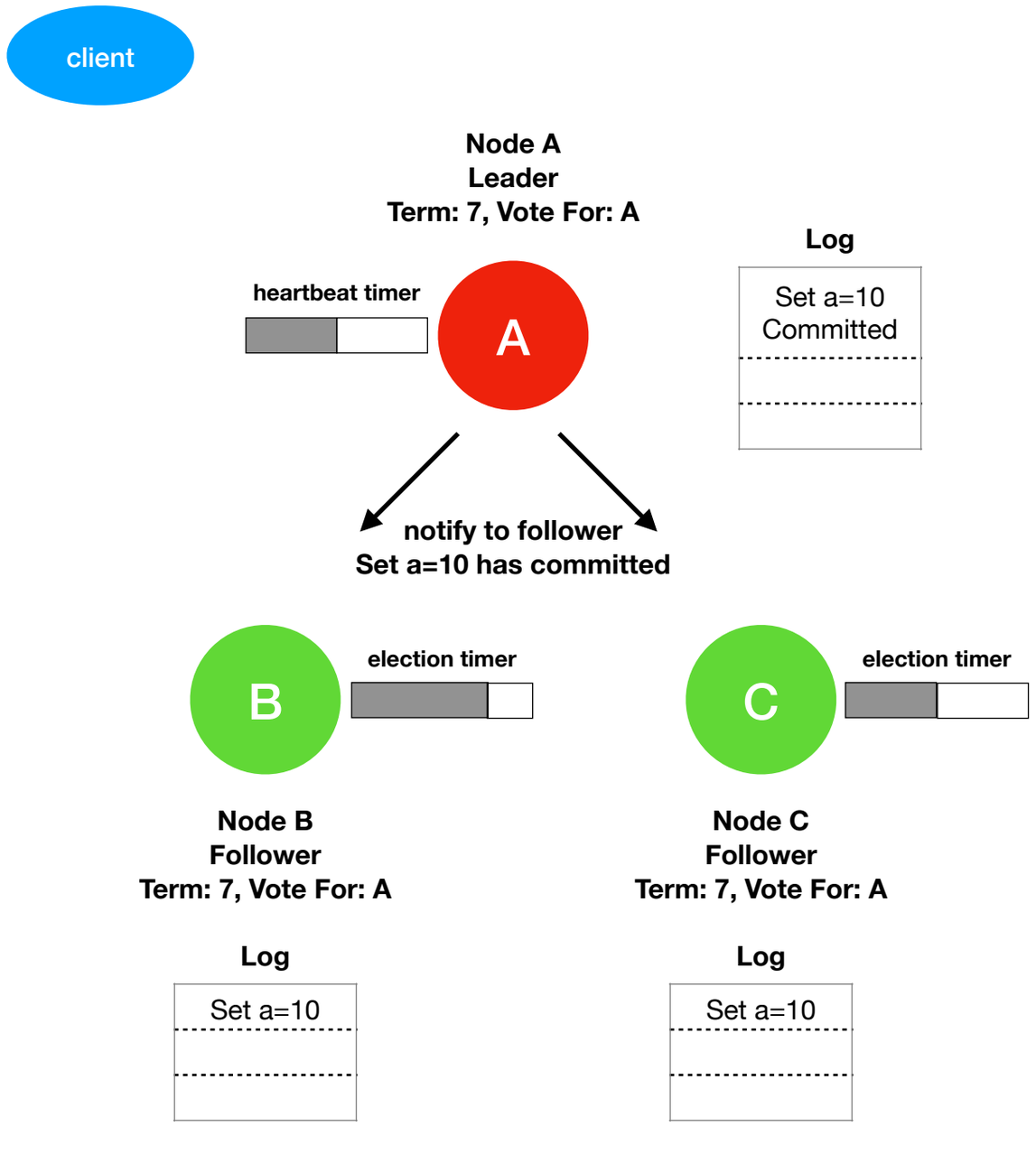
日志复制

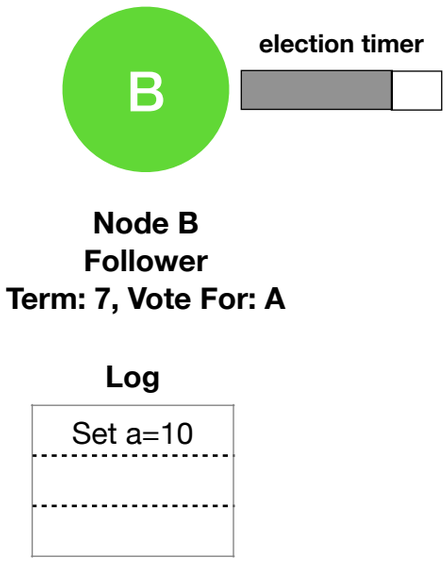


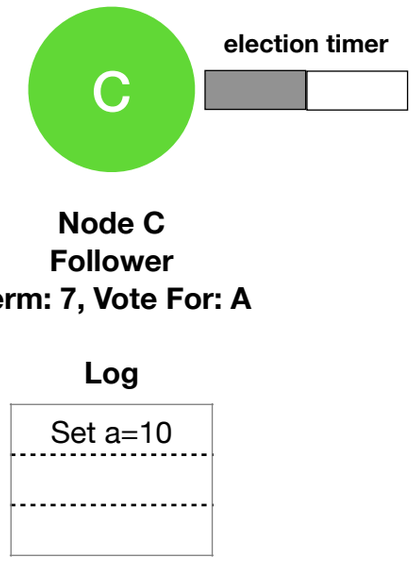
Log Replication Process

日志复制









Log Replication Process

日志复制

Each Node:

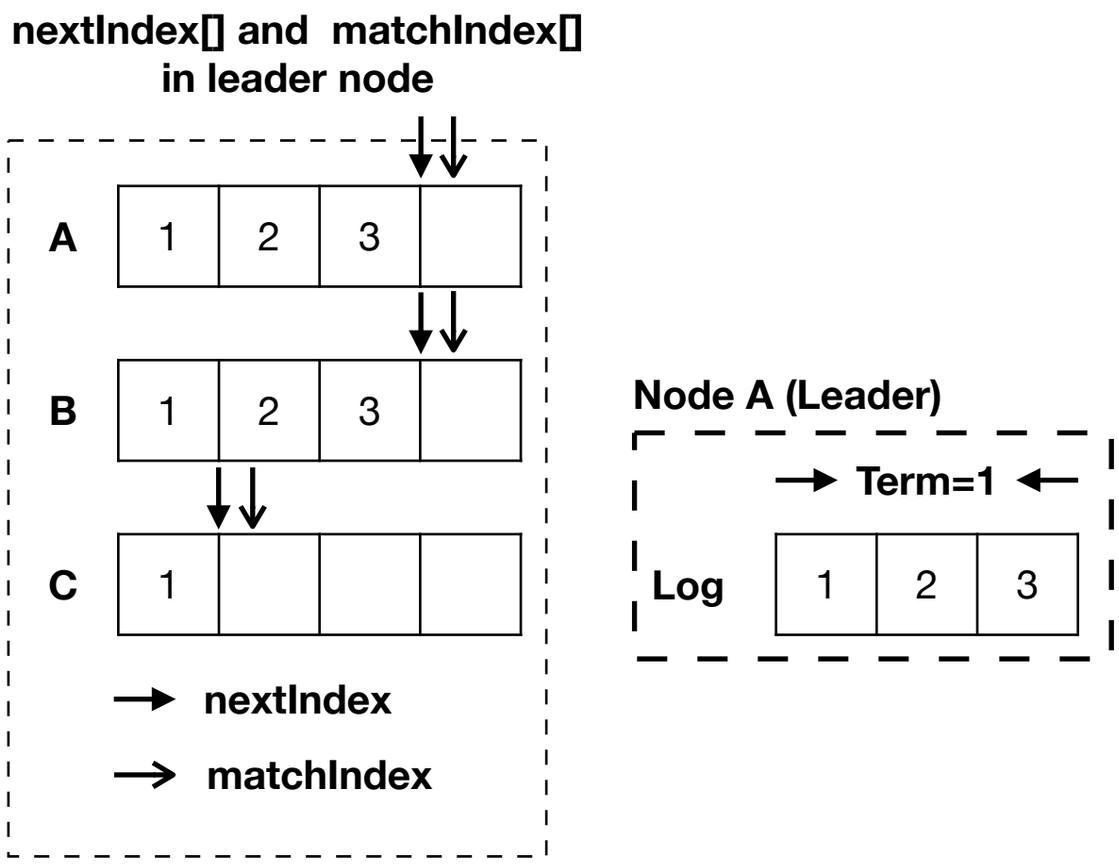
- **commitIndex: the max log index which has been committed**
- **lastApplied: the max log index which has been apply to state machine**

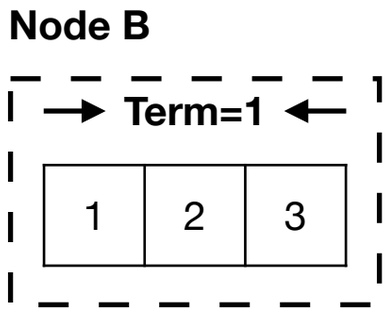
Leader Node:

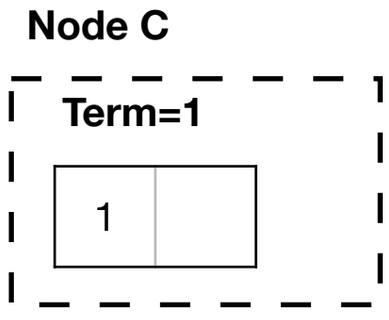
- **nextIndex[]: log index need to send to follower nodes**
- **matchIndex[]: max log index has been sent to follower nodes**

Data Struct

日志复制

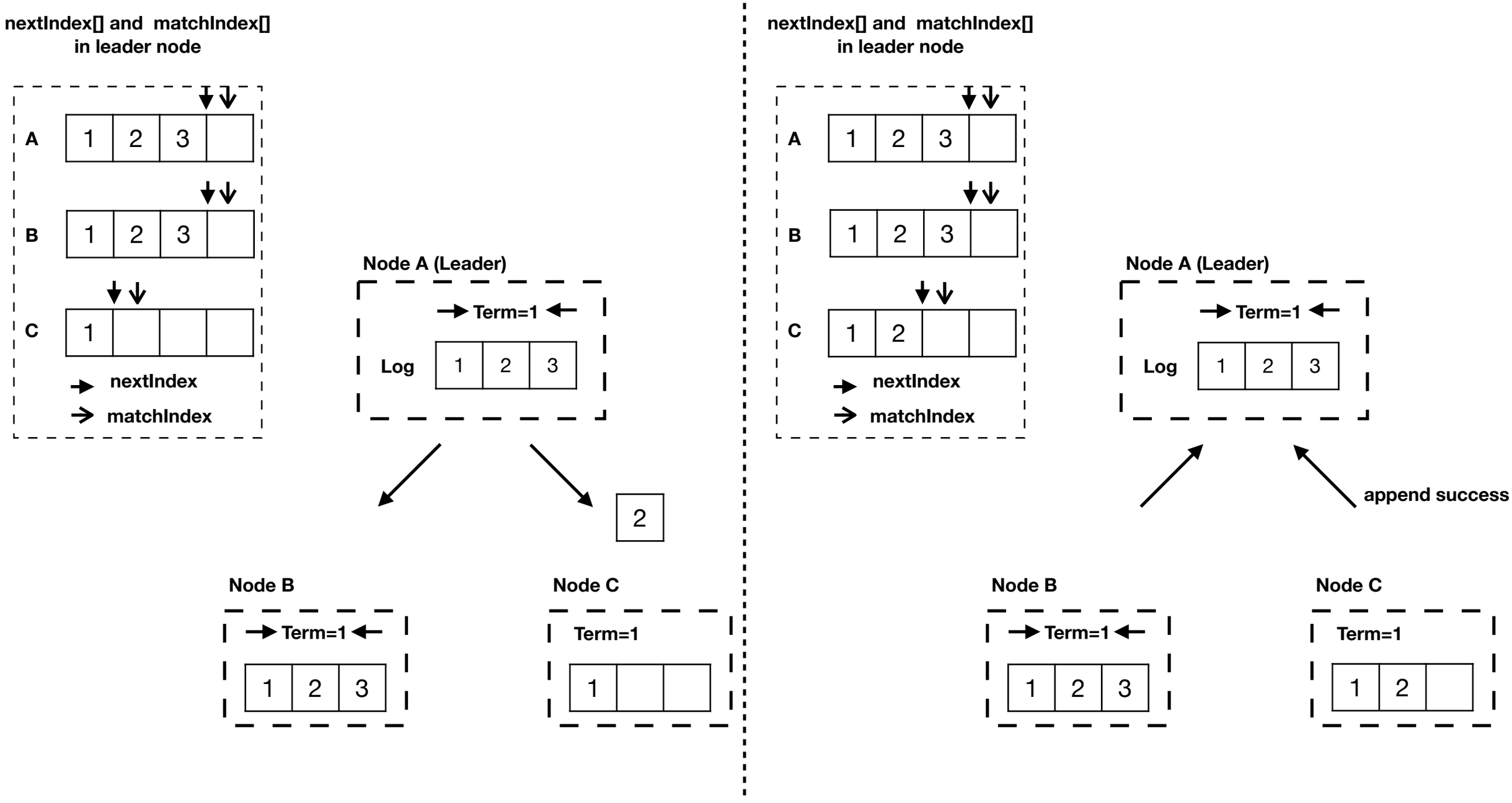






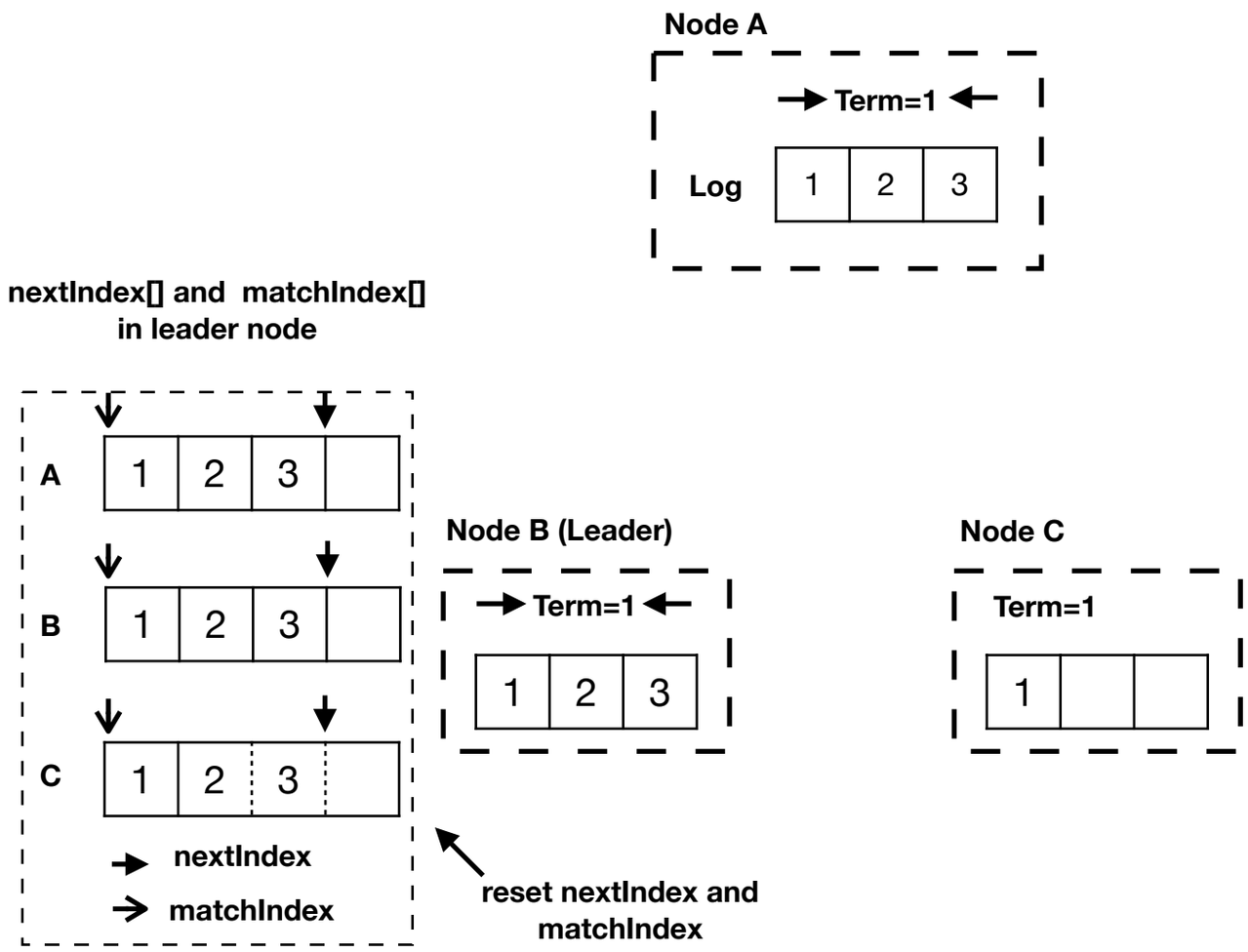
Log Replication

日志复制



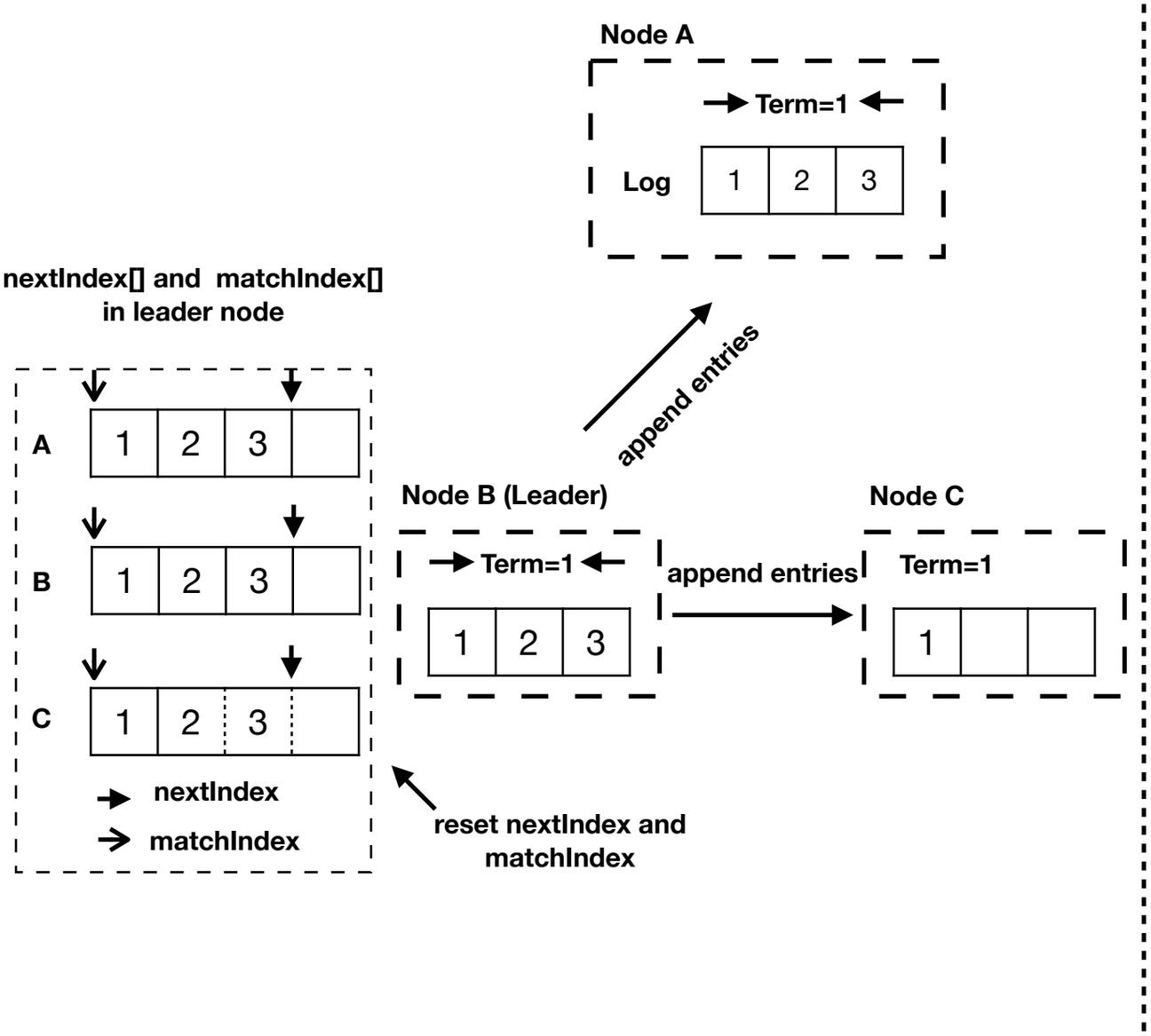
Log Replication

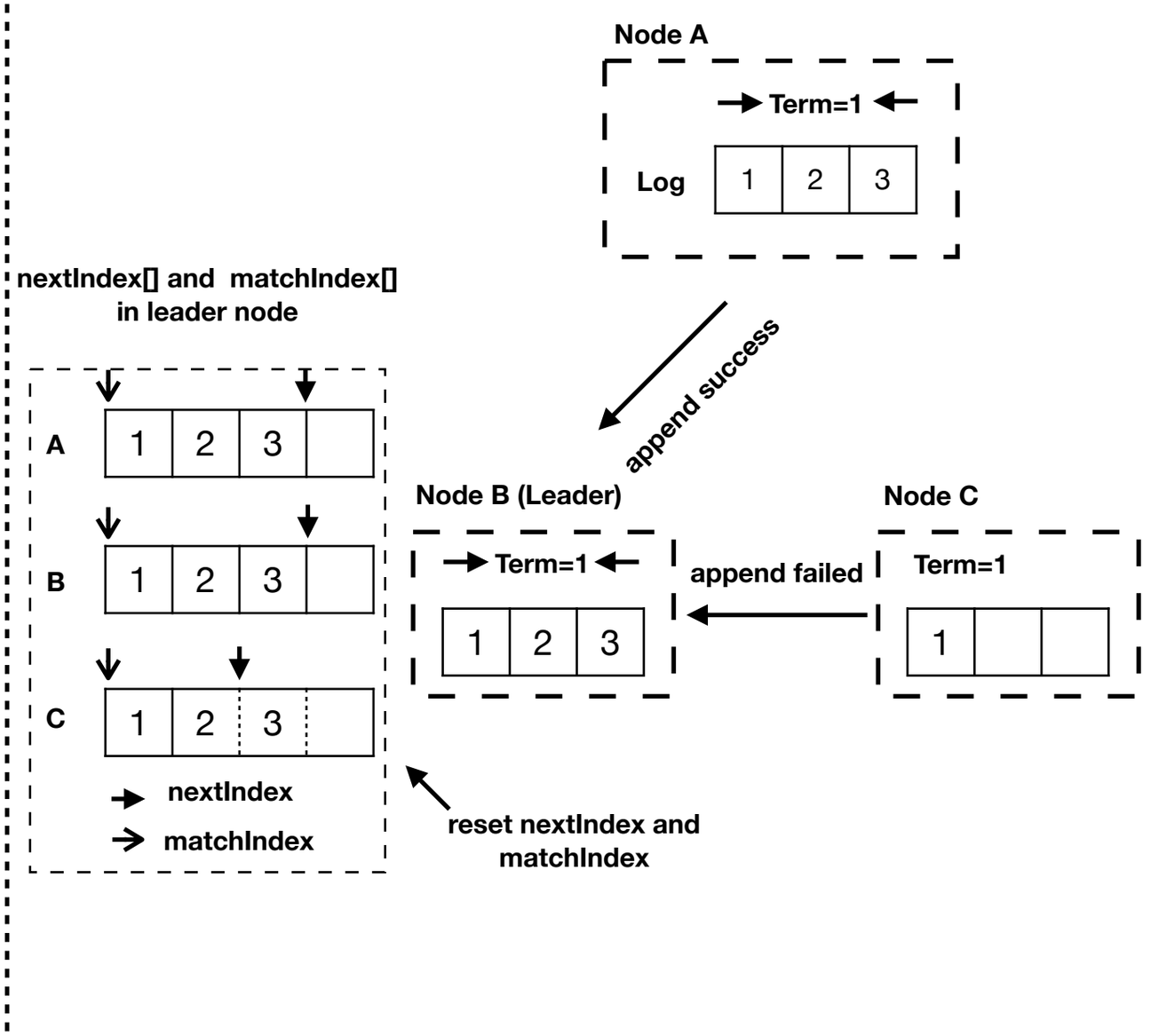
日志复制



Log Replication After Leader Crash

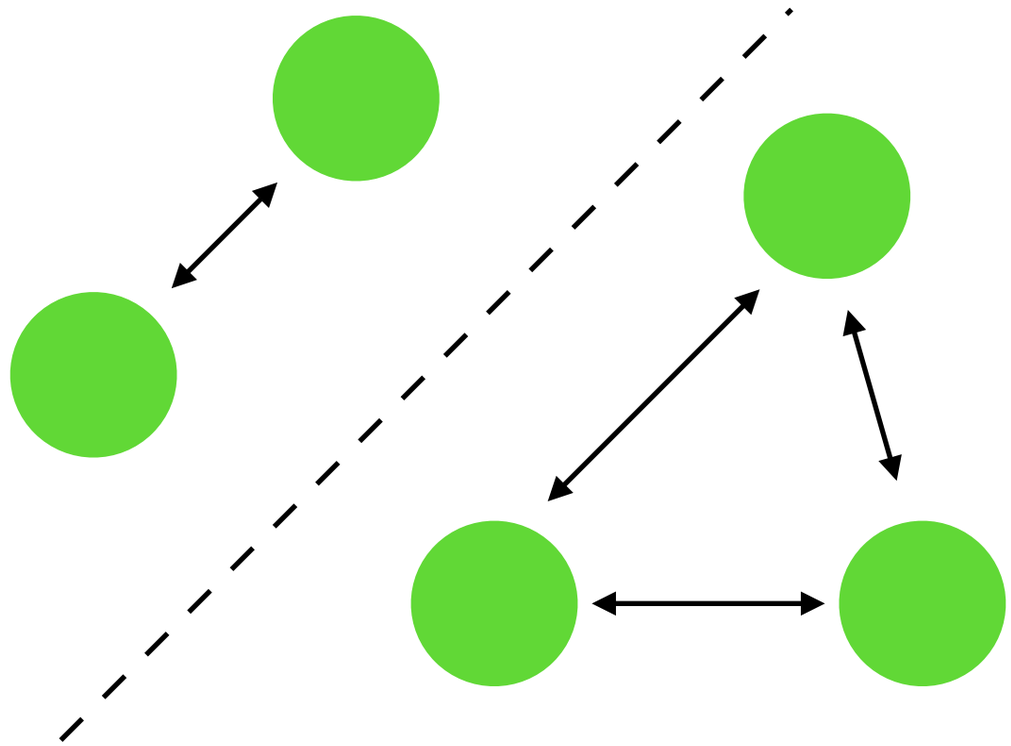
日志复制





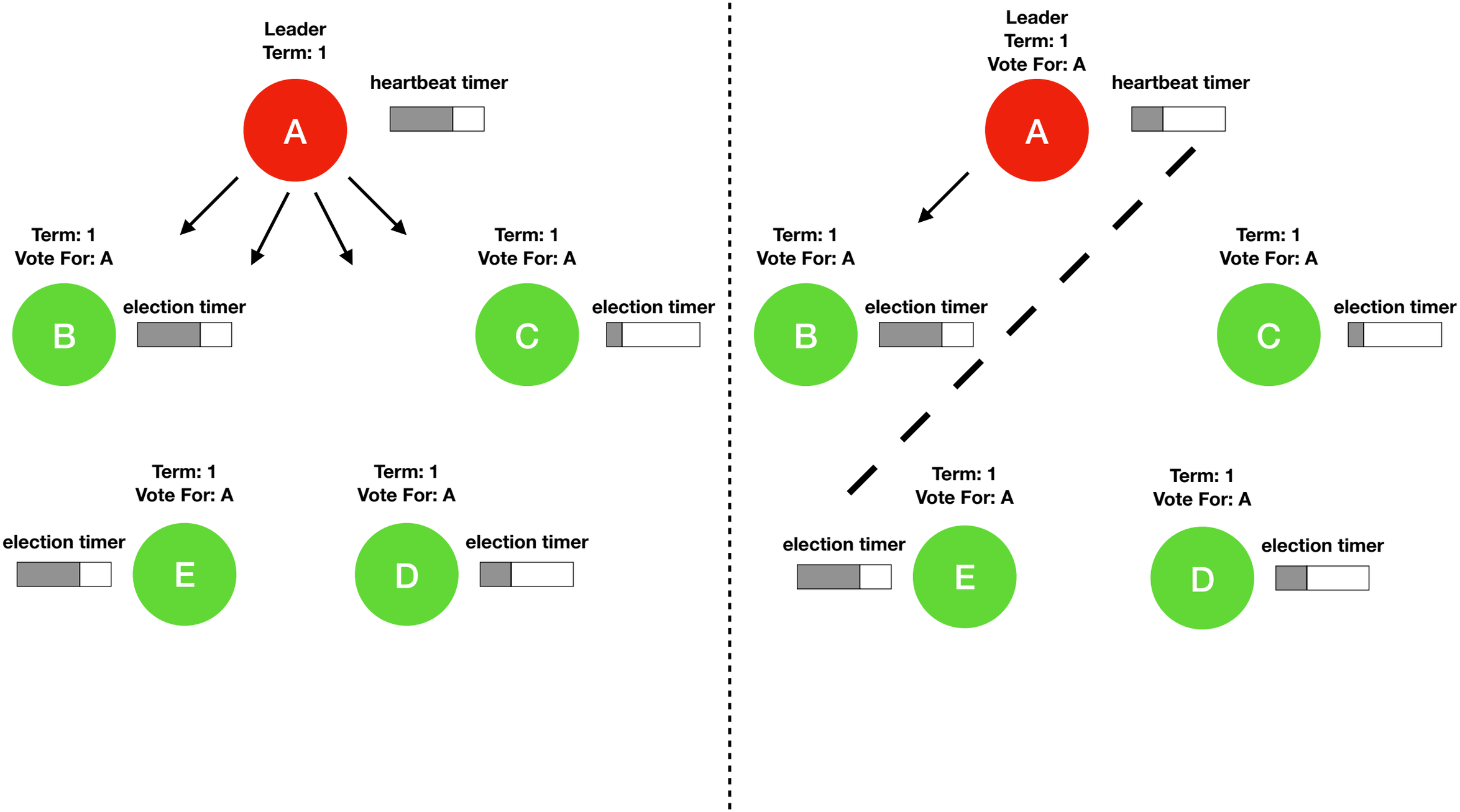
Log Replication After Leader Crash

网络分区场景



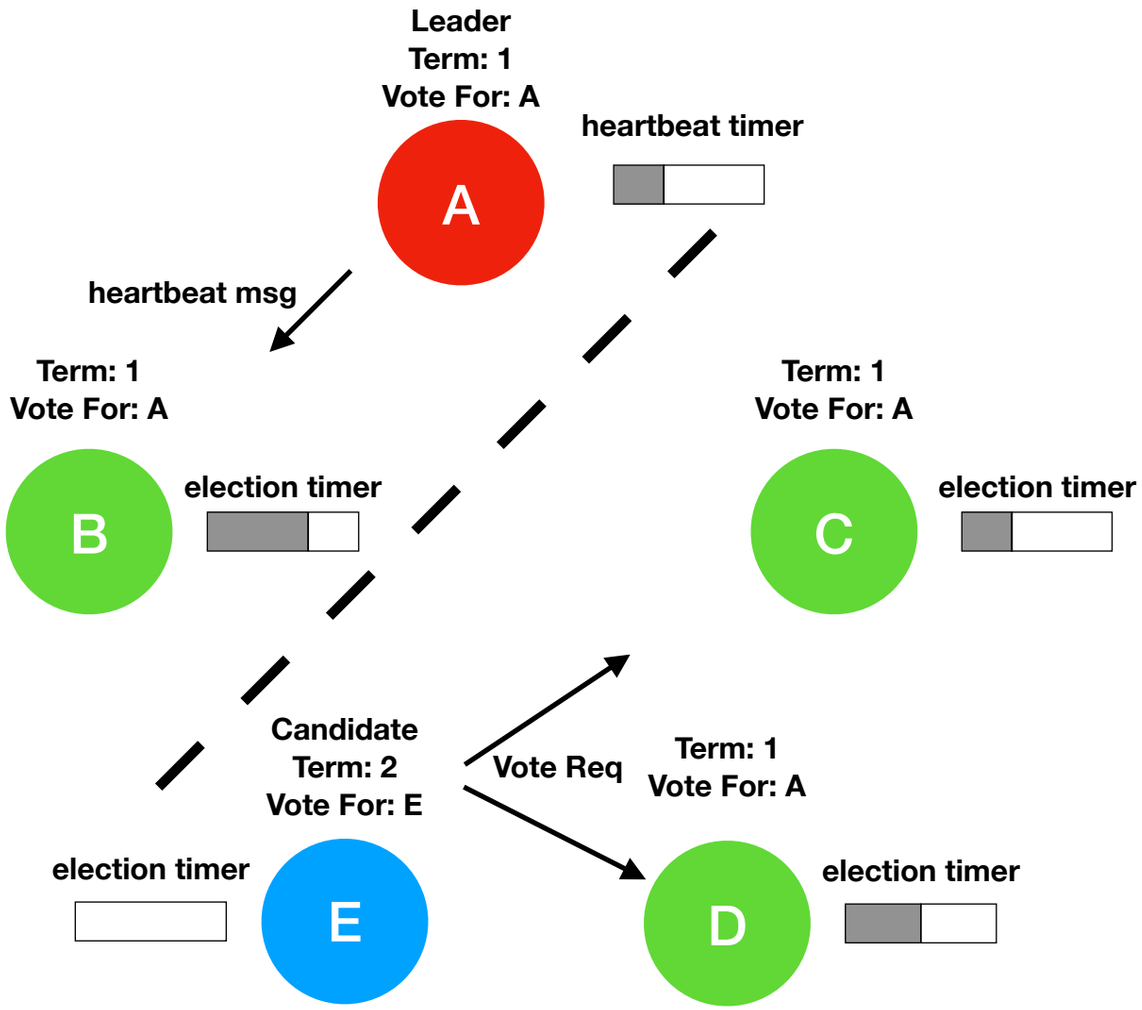
Network Isolation

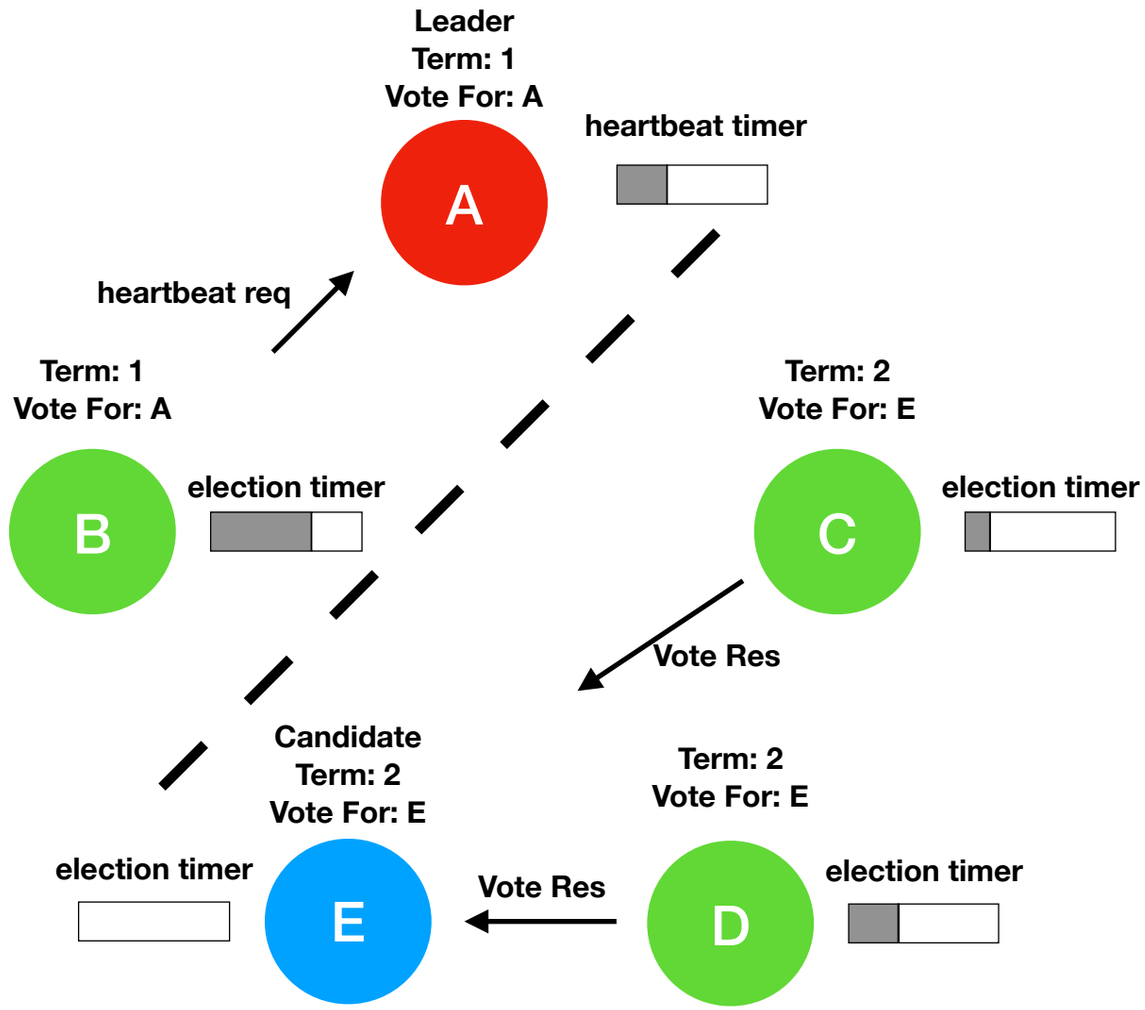
网络分区场景



Network Isolation

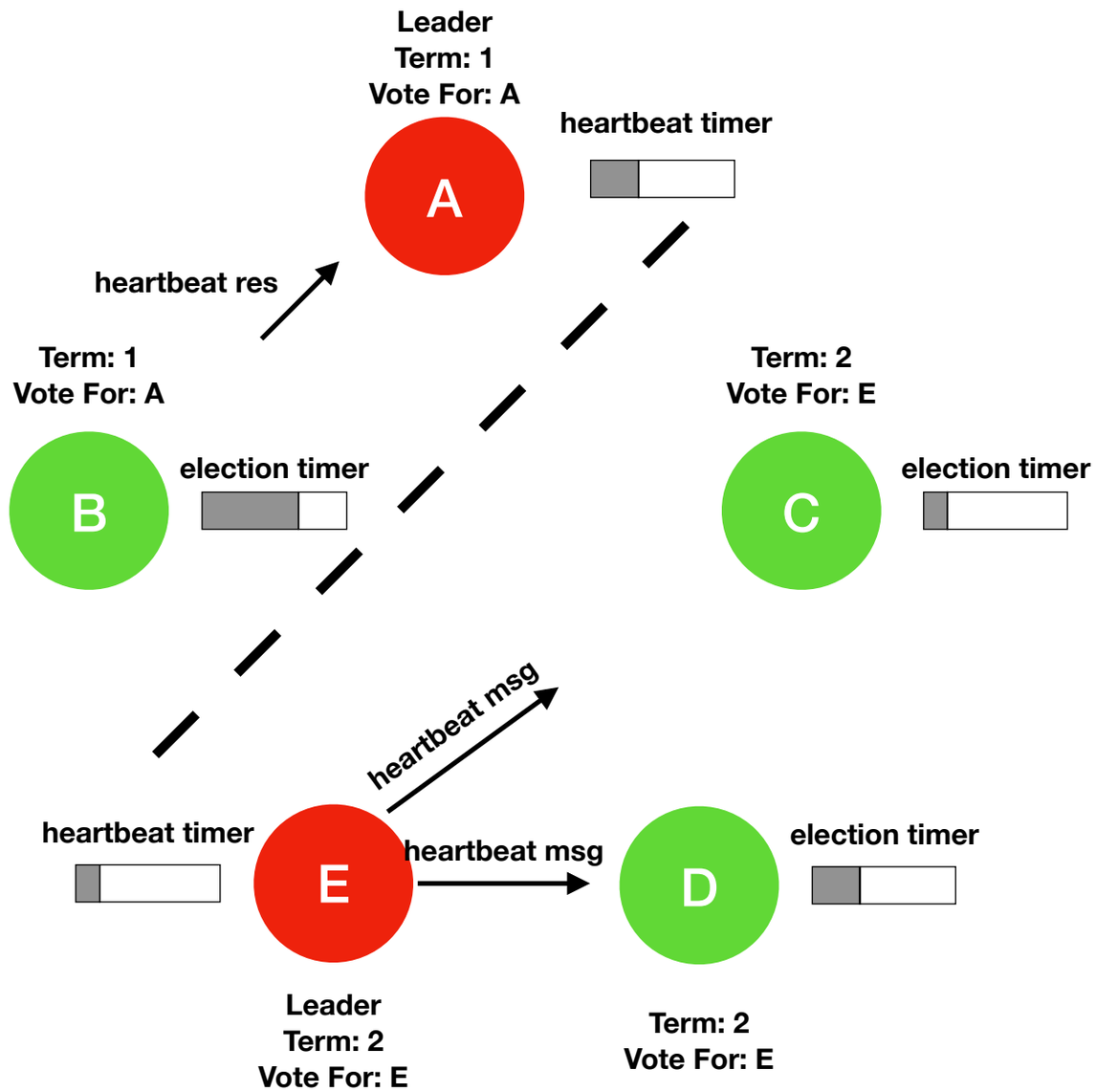
网络分区场景





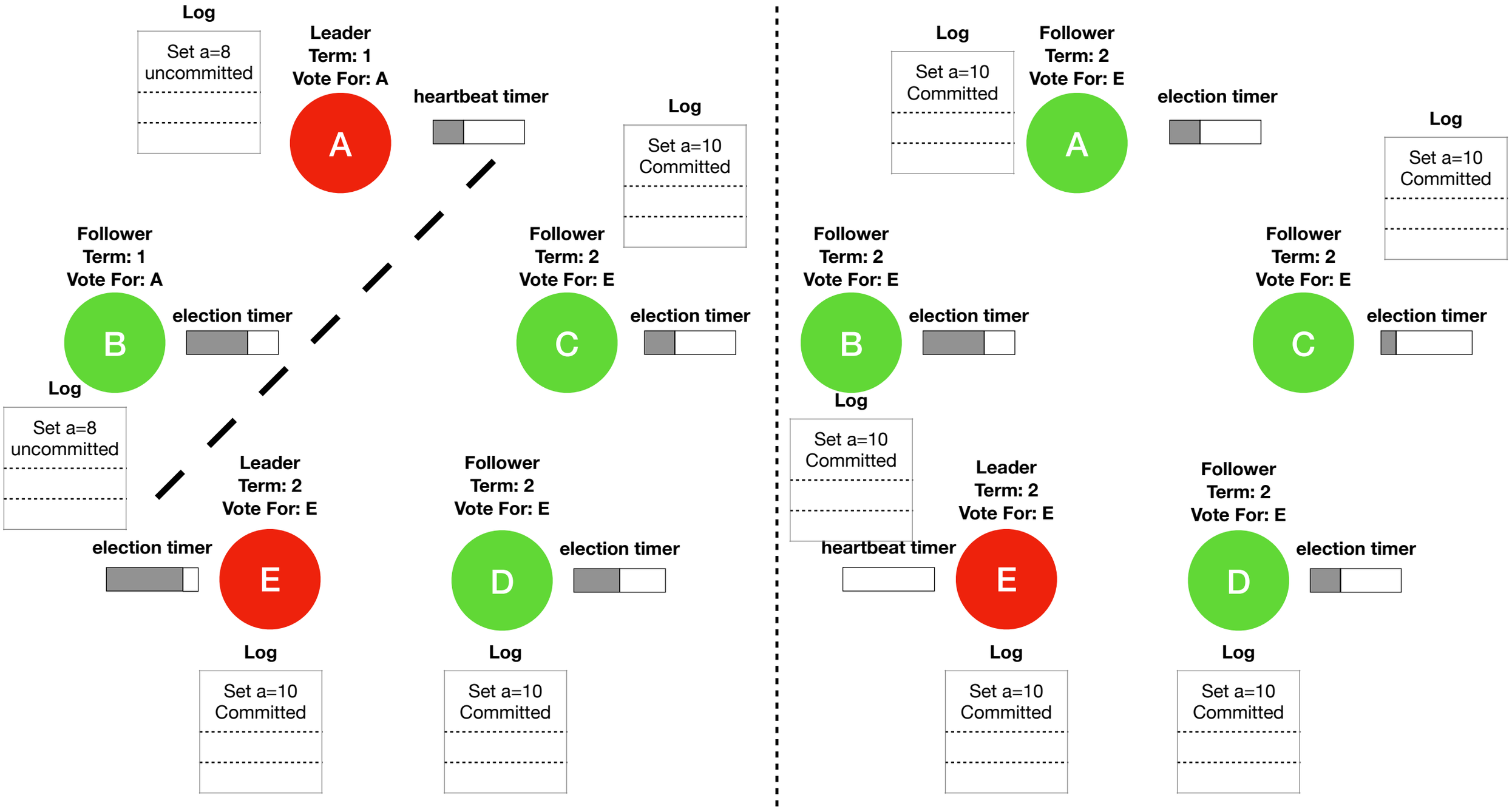
Network Isolation

网络分区场景



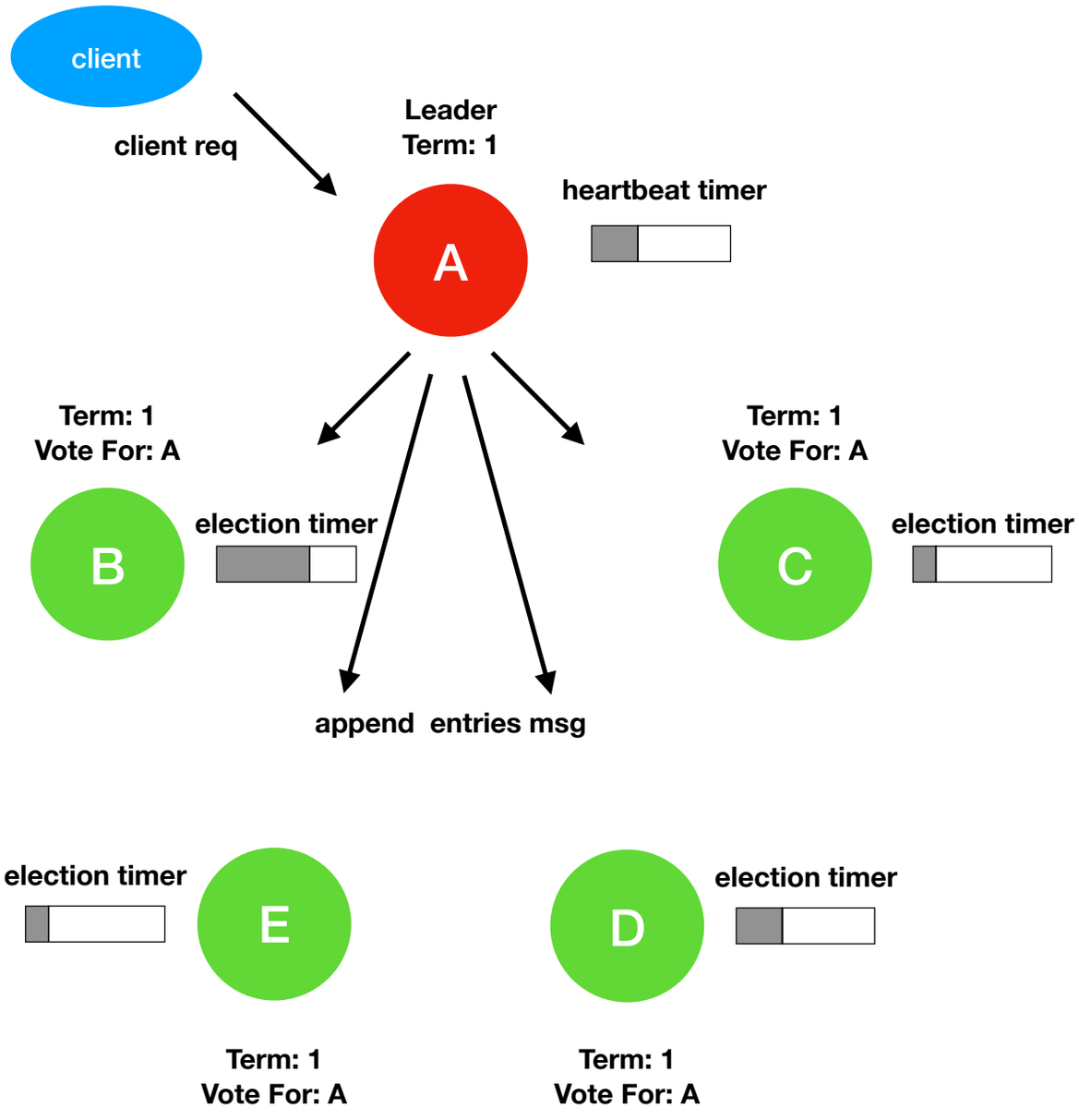
Network Isolation

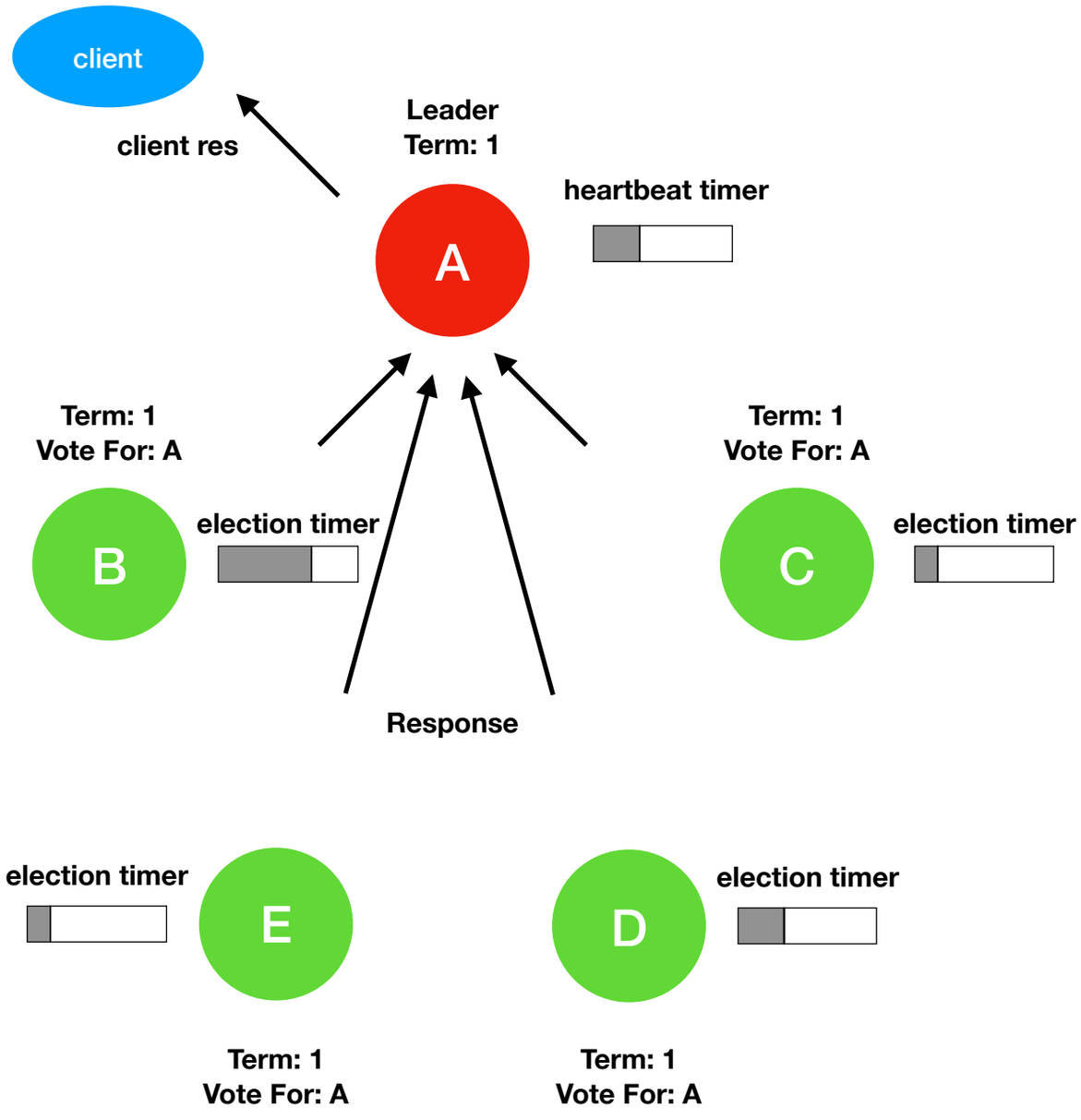
网络分区场景



Network Isolation

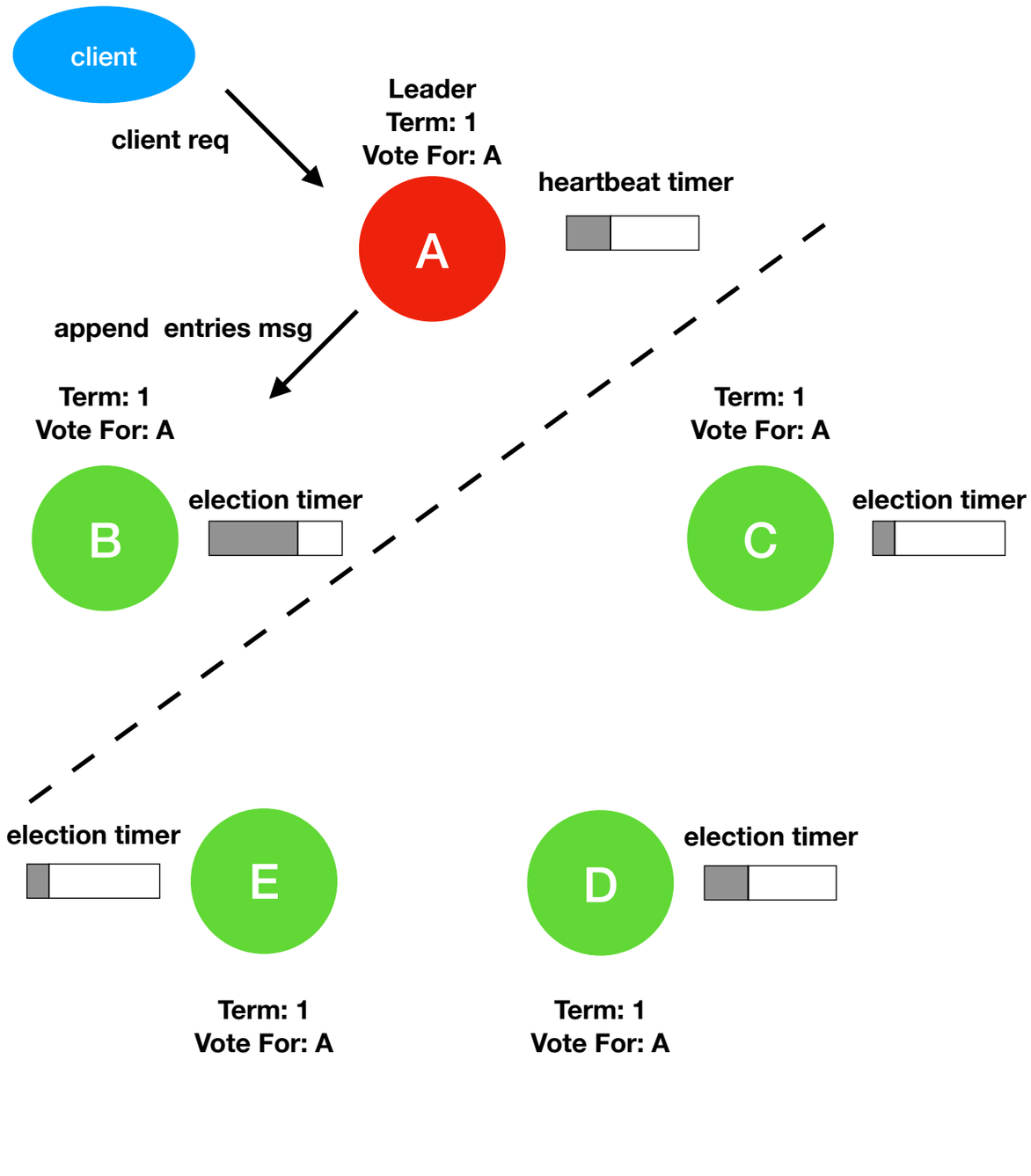
网络分区场景

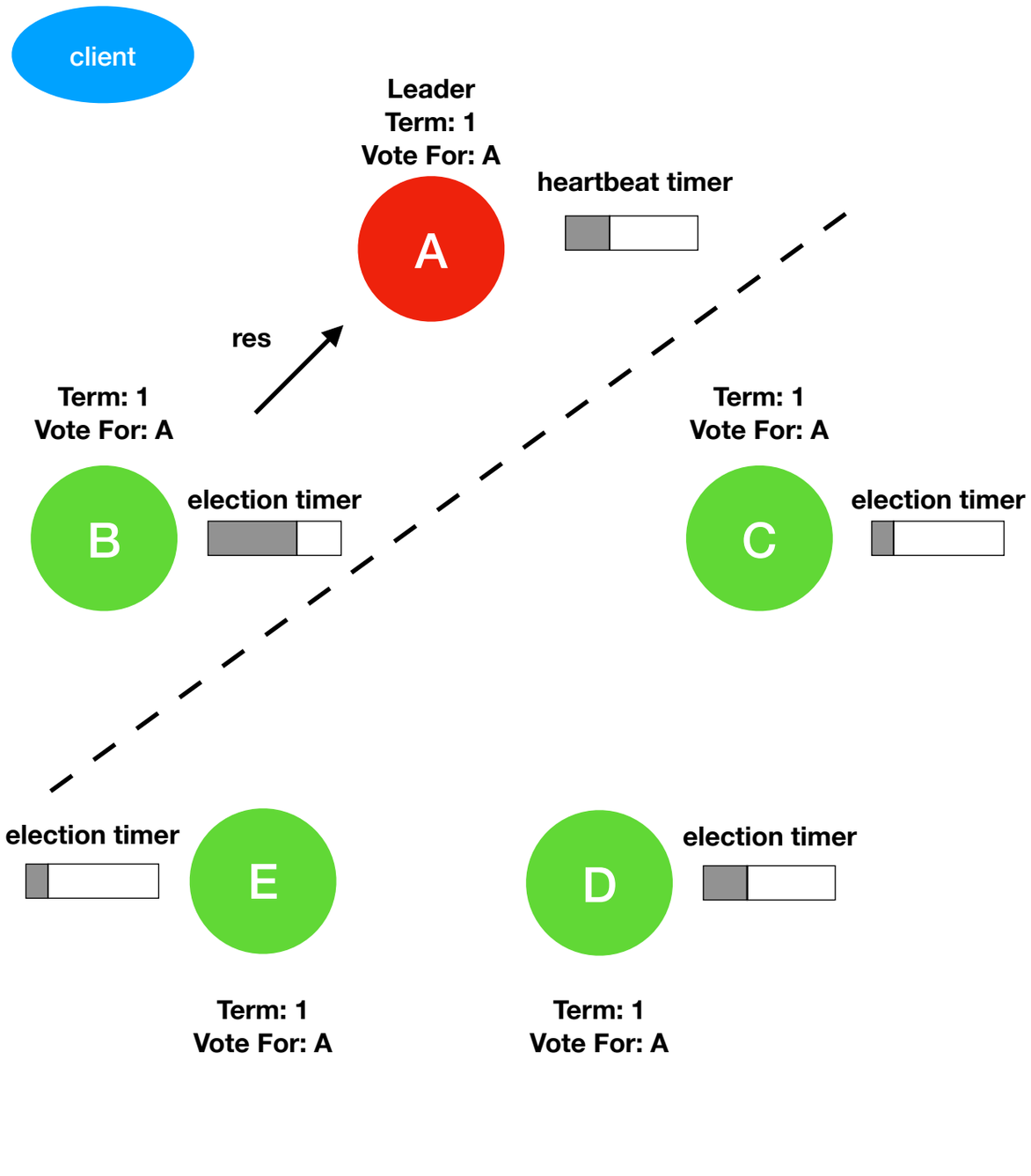




Client Reques

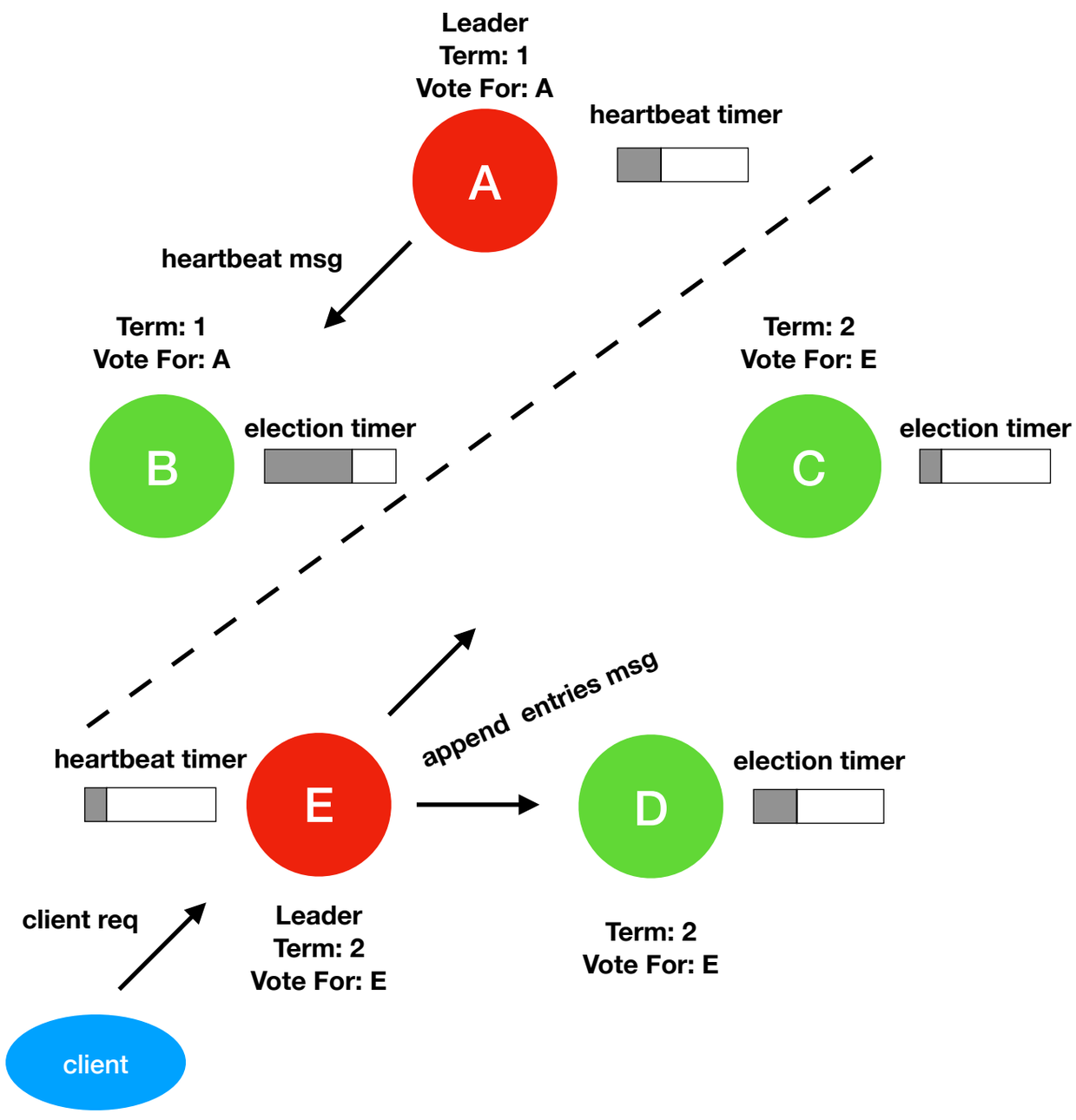
网络分区场景

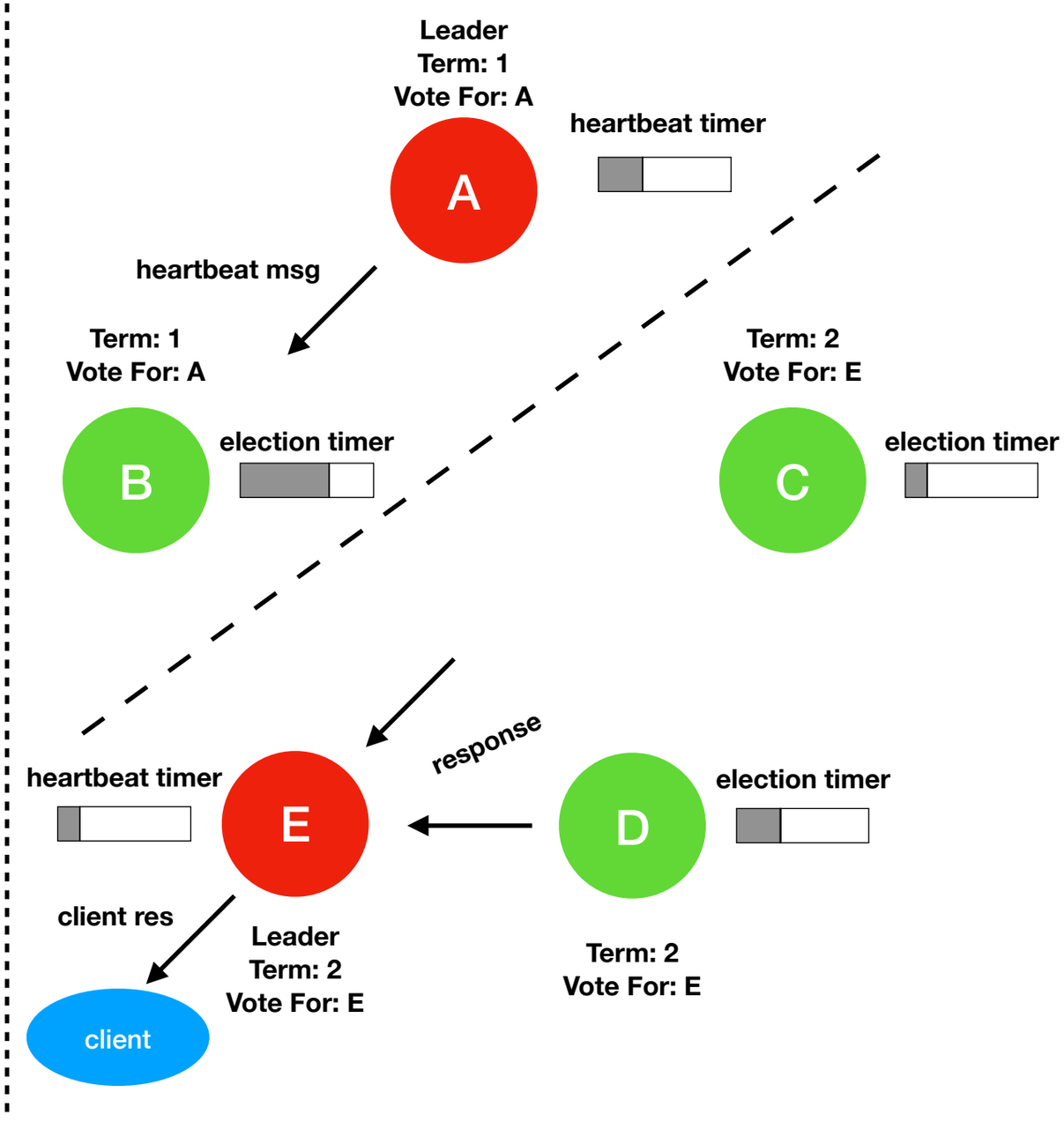




Client Request

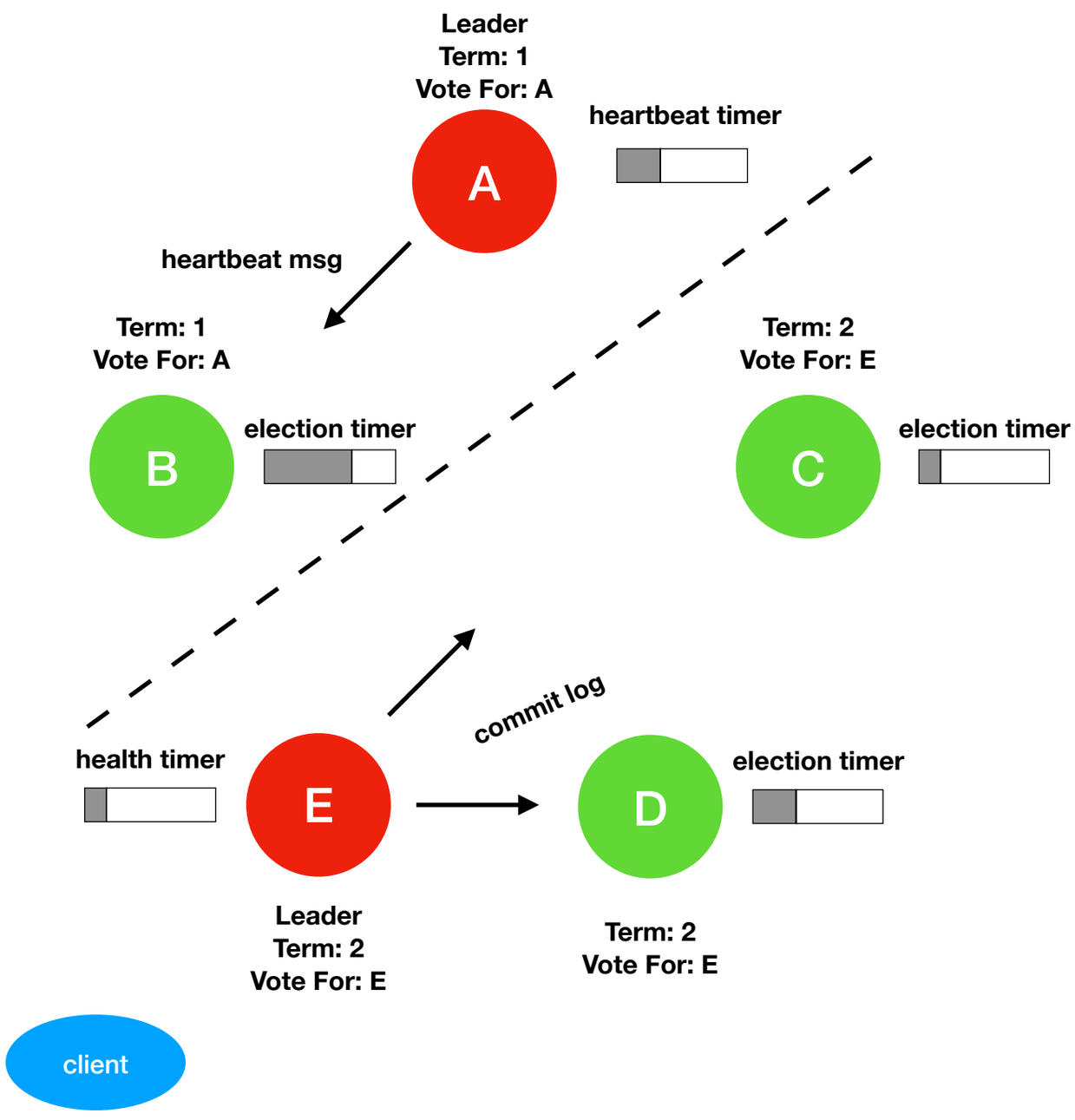
网络分区场景

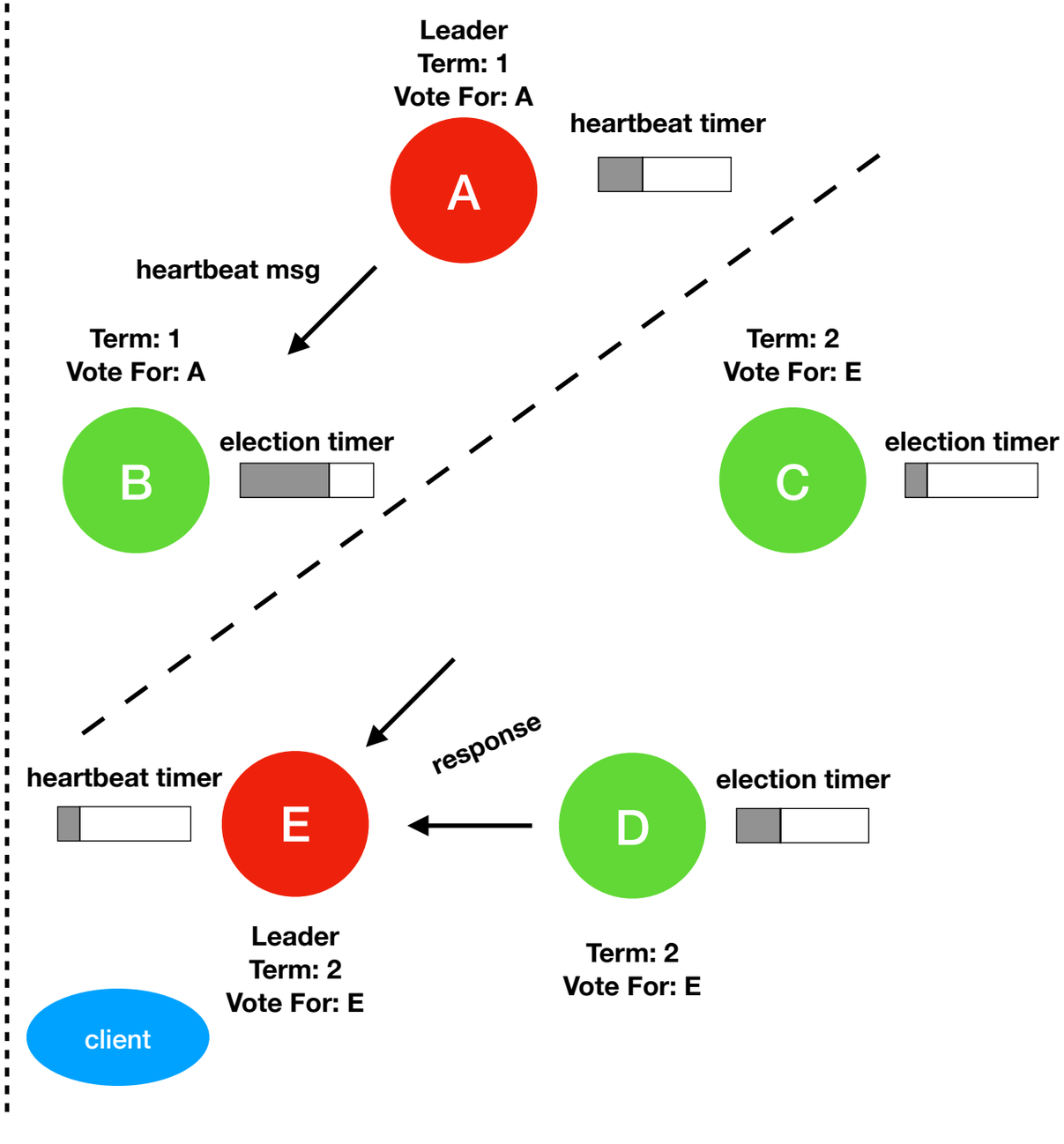




Client Request

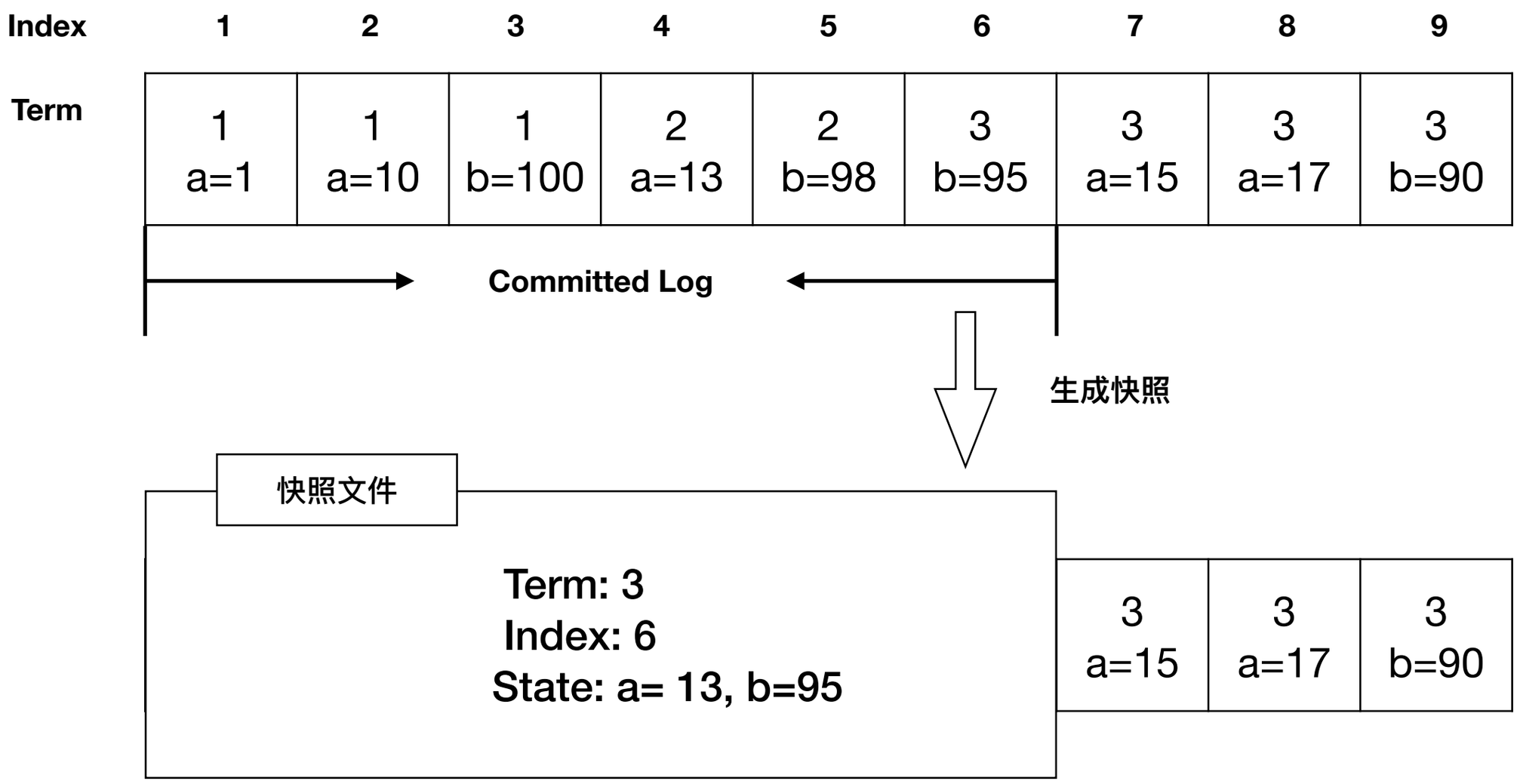
网络分区场景





Client Request

日志压缩与快照



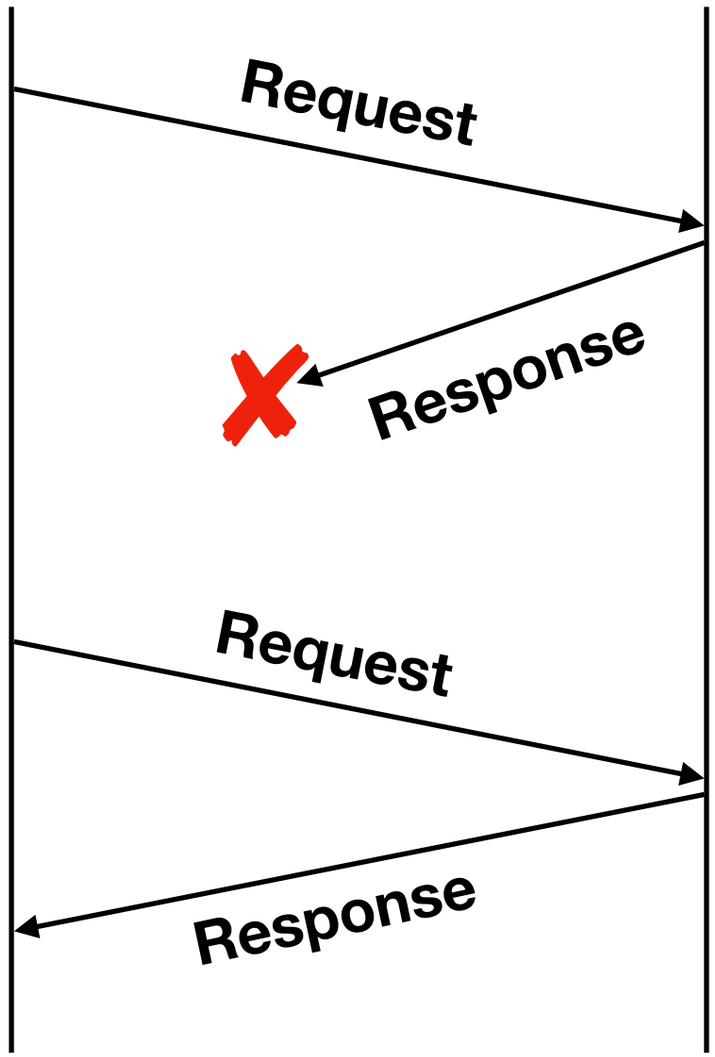
Snapshot Struct

其它技术点

- linearizable 语义
- 只读请求
- PreVote状态
- Leader节点转移

Other Points

其它技术点



linearizable 语义

其它技术点

- Leader节点在其任期开始时提交一条空日志记录，保证上一个任期中的所有日志都会被提交

- Leader节点会记录该只读请求对应的编号作为readIndex ， 当Leader节点的提交位置(commitIndex)达到或是超过该位置之后，即可响应该只读请求

- Leader节点在处理只读的请求之前必须检查集群中是否有新的Leader 节点(PreVote)，必须由新Leader 节点来处理此次只读请求

- 随着日志记录的不断提交， Leader 节点的提交位置（ commitIndex ） 最终会超过上述readIndex ，此时Leader 就可以响应客户端的只读请求了

Read-only Request

其它技术点

当某个节点要发起选举之前，需要先进入PreVote 的状态。在PreVot已状态下的节点会先尝试连接集群中的其他节点，如果能够成功连接到半数以上的节点，才能真正切换成Candidate 状态并发起新一轮的选举。

PreVote State

其它技术点

- 暂停接收客户端请求
- 让一个指定的Follower节点的本地日志与当前的Leader节点完全同步
- 该特定的Follower节点立刻发起新一轮的选举

Leader Node Transfer

The End