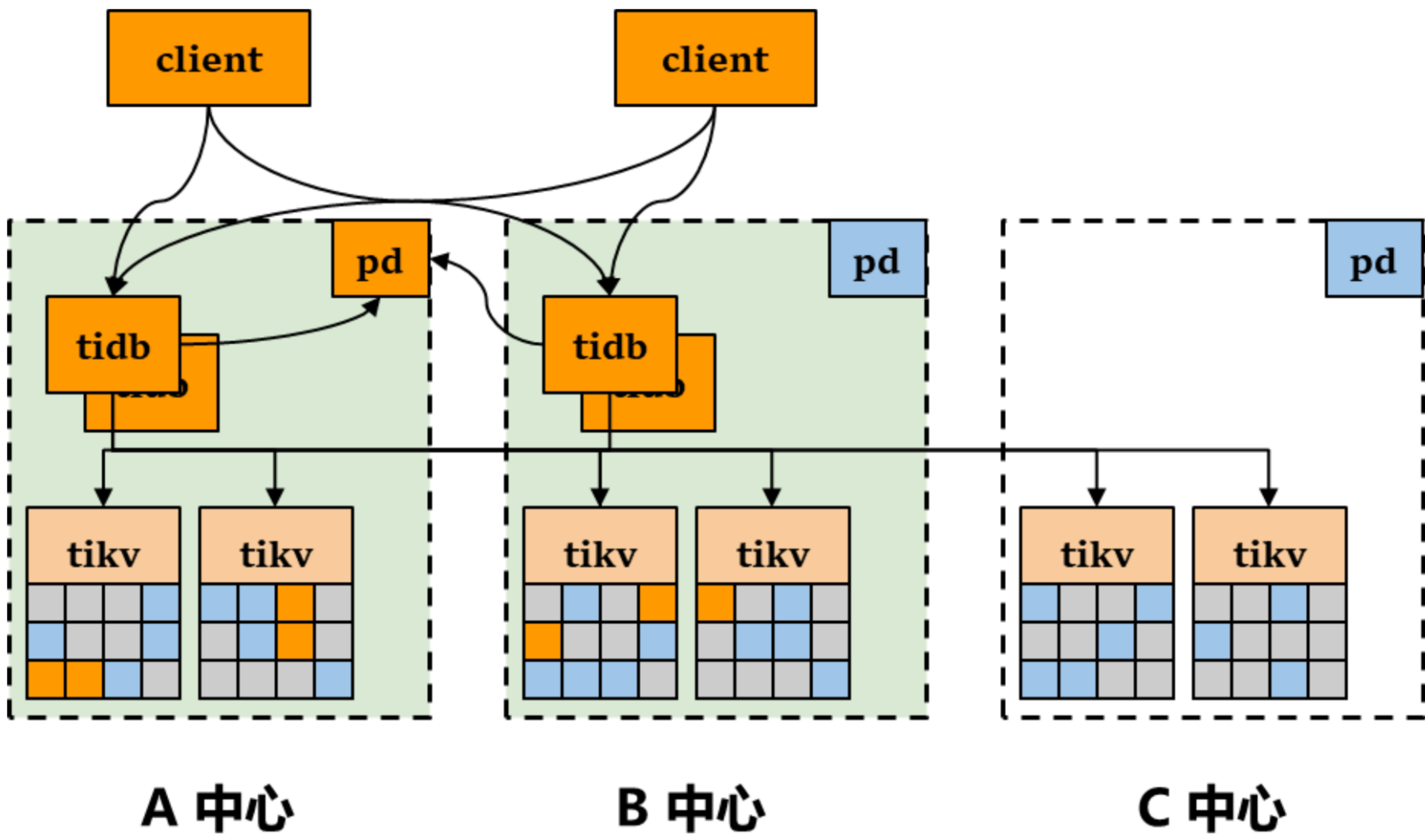


## TiDB 三中心"脑裂"场景探讨

K Kassadar 发表于 2024-03-22

原创

TiDB 以其卓越的高可用性而闻名。然而，在跨多个数据中心进行部署时，数据中心之间的特殊网络拓扑可能带来额外的可用性挑战。让我们详细分析在多数据中心部署环境中，特定的网络拓扑是如何对系统的整体可用性产生影响的。

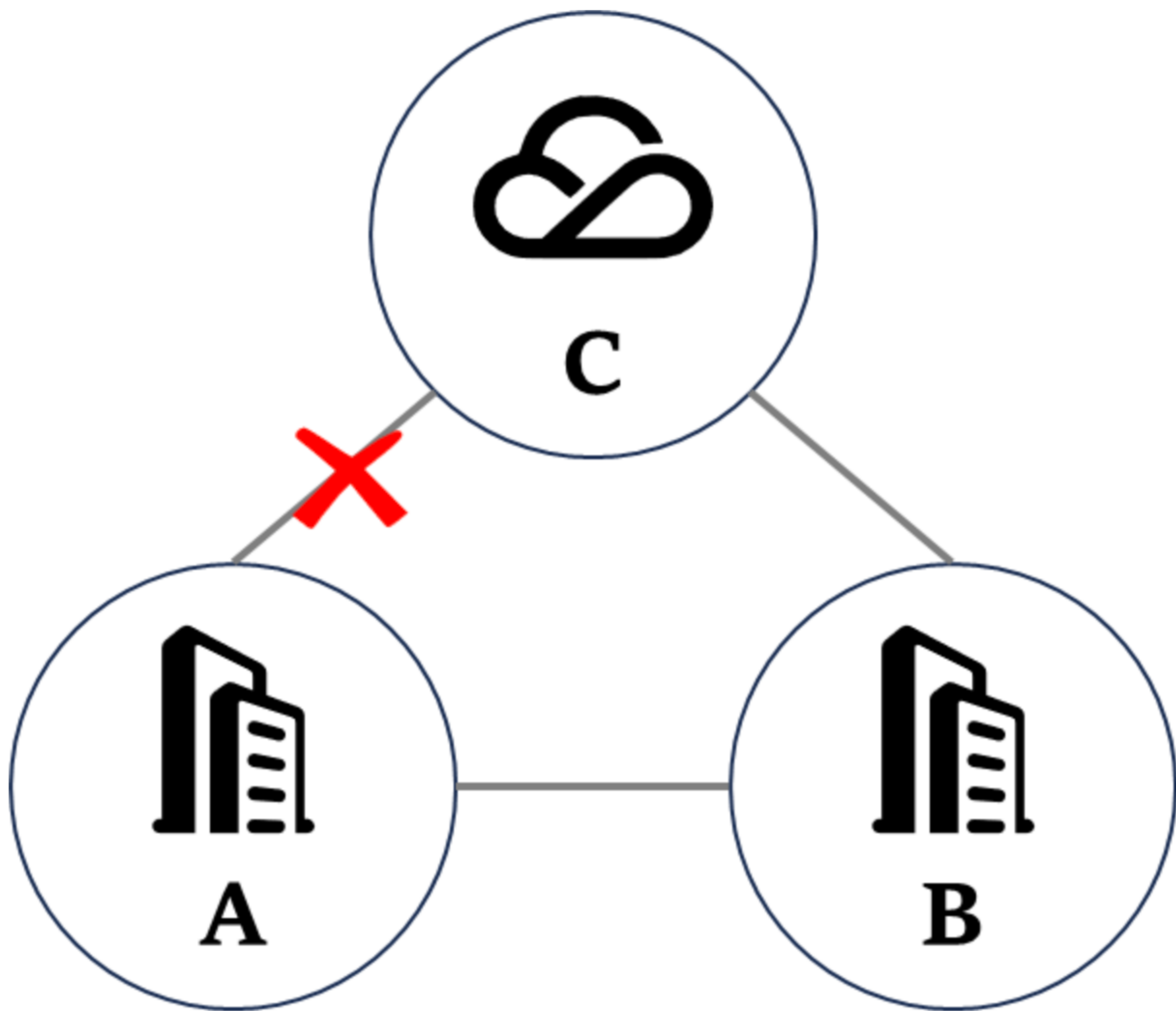


部署架构简述：

- 云下 A、B 两中心提供服务，两中心对等，无服务优先级顺序
- 云上 C 中心不提供服务，但存储数据，作用类似仲裁节点
- Region Leader 优先位于 A、B 两中心之内
- Pd Leader 优先位于 A、B 两中心之内
- 应用仅部署在 A、B 机房中，应用访问不考虑本地亲和性

具体配置可参照 [双区域多 AZ 部署 TiDB](#)

### 场景一



第一阶段：A 机房和 C 机房间网络断开（半断开 C）

预期情况：A、B 机房均可提供完整的服务

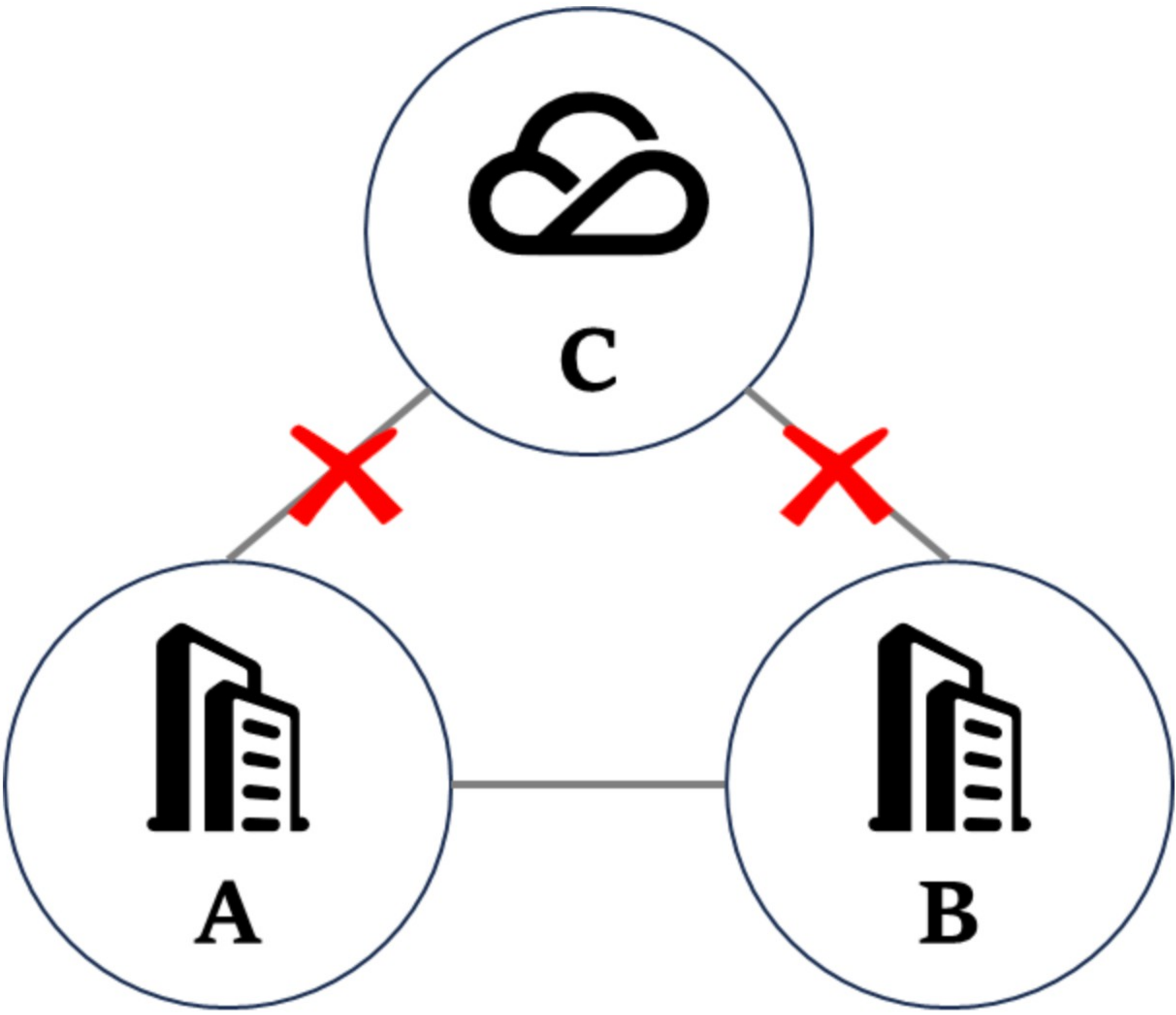
说明：读写由 A、B 机房同时提供，C 机房不提供服务，A 机房和 C 机房之间网络断开后：

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 B 机房组成子区域，满足多数派
  - B 机房有 Leader：B 机房仍然可以和 A、C 机房组成子区域，满足多数派
- 从应用角度看：



- A 机房可访问 A、B 机房的数据
- B 机房可访问 B、A 机房的数据

因此 A、B 机房均可提供完整的服务



第二阶段：在 1 的基础上增加 B 机房和 C 机房间网络断开（完全断开 C）。

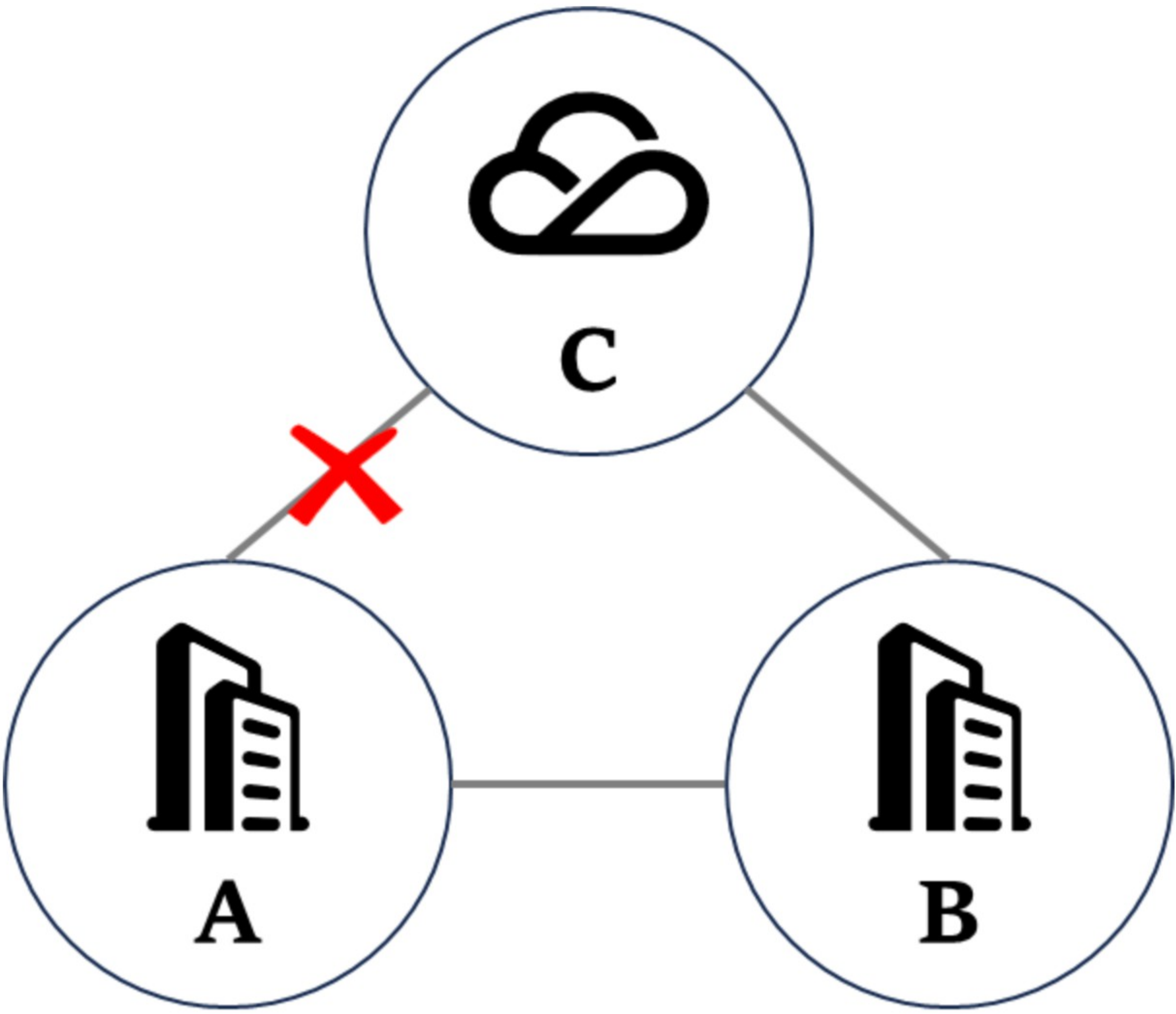
预期情况：A、B 机房均可提供完整的服务

说明：读写由 A、B 机房同时提供，C 机房被完全隔离：

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 B 机房组成子区域，满足多数派
  - B 机房有 Leader：B 机房仍然可以和 A 机房组成子区域，满足多数派
- 从应用角度看：
  - A 机房可访问 A、B 机房的数据
  - B 机房可访问 B、A 机房的数据

因此 A、B 机房均可提供完整的服务

## 场景二



第一阶段：调整 C 机房可承载 Region Leader；A 机房和 C 机房间网络断开（半断开 C）

预期情况：A 机房不可提供完整的服务，B 机房可提供完整的服务

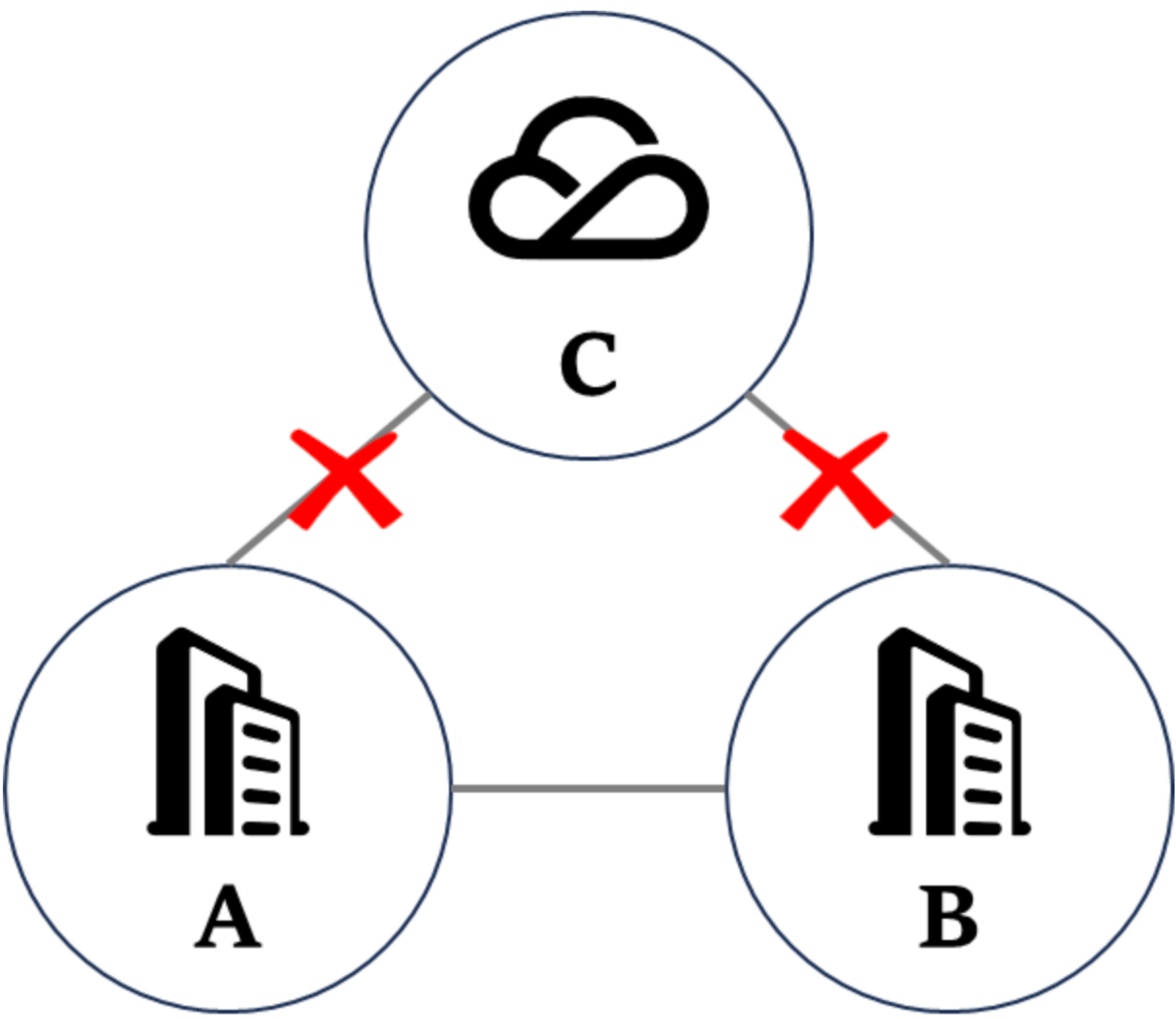
说明：读写由 A、B、C 机房同时提供，A 机房到 C 机房网络断开后：

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 B 机房组成子区域，满足多数派
  - B 机房有 Leader：B 机房仍然可以和 A、C 机房组成子区域，满足多数派
  - C 机房有 Leader：C 机房仍然可以和 B 机房组成子区域，满足多数派
- 从应用角度看：



- A 机房可访问 B 机房的数据，不能访问 C 机房的数据
- B 机房可访问 A、C 机房的数据

因此，B 机房可提供完整的服务；A 机房无法提供完整的服务，其访问 B 时正常，而其访问 C 时会报错。



第二阶段：B 机房和 C 机房间网络断开（完全断开 C）

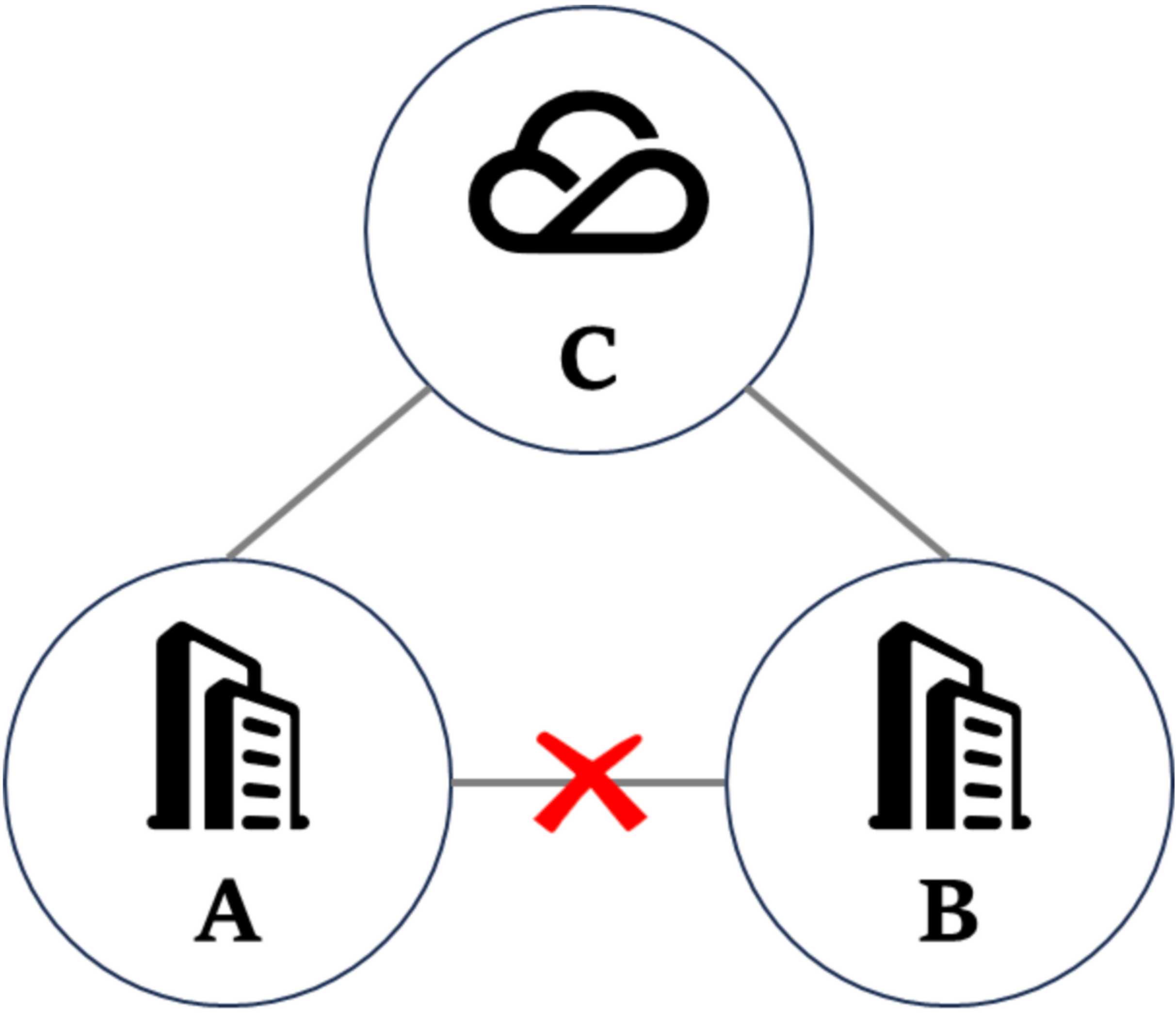
预期情况：A、B 机房均可提供完整的服务

说明：读写由 A、B、C 机房同时提供，C 被完全隔离

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 B 机房组成子区域，满足多数派
  - B 机房有 Leader：B 机房仍然可以和 A 机房组成子区域，满足多数派
  - C 机房没有 Leader
- 从应用角度看：
  - A 机房可访问 B 机房的数据
  - B 机房可访问 A 机房的数据

因此 A、B 机房均可提供完整的服务

### 场景三



第一阶段：A 机房和 B 机房间网络断开（半断开 B）

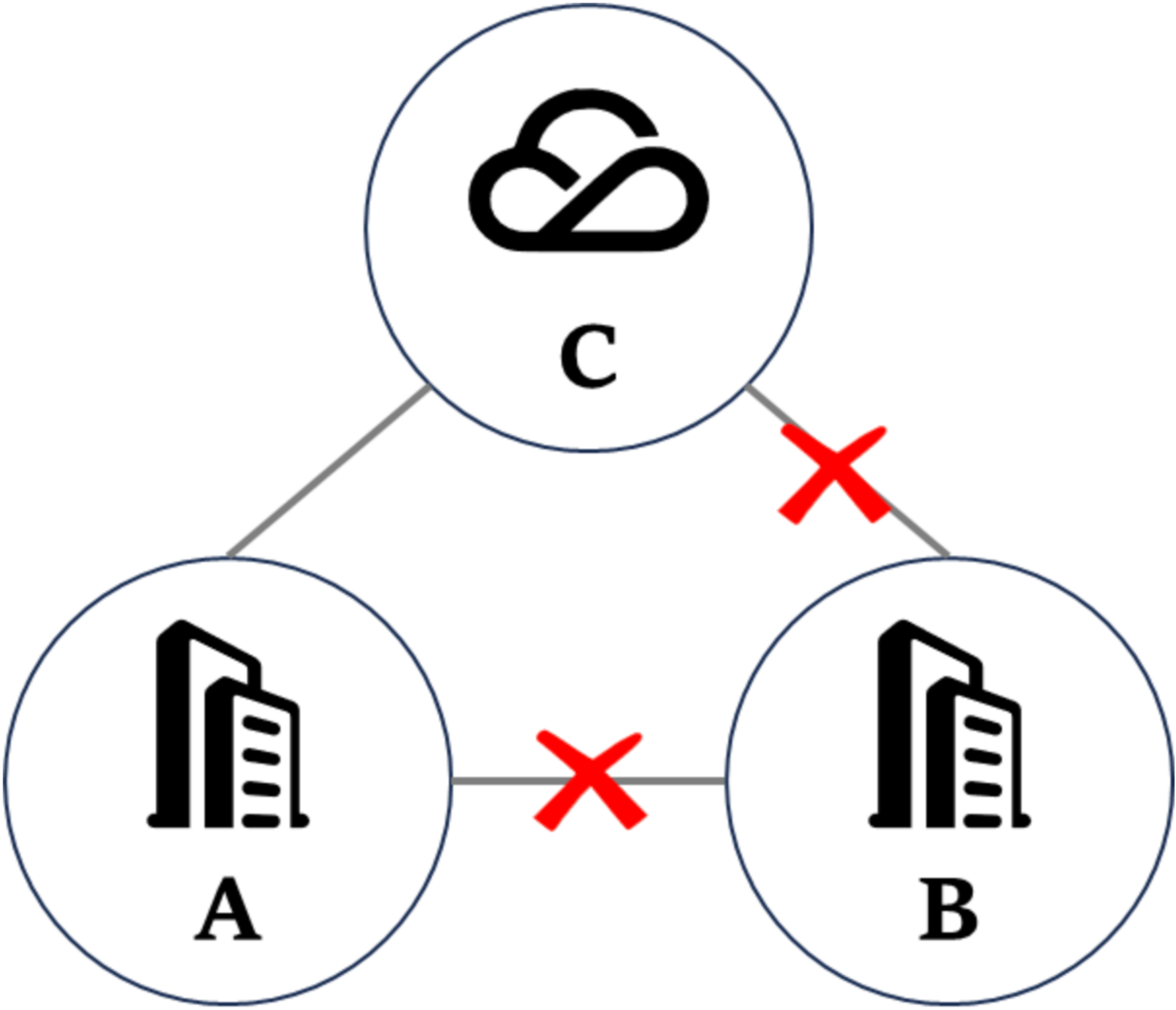
预期情况：A、B 机房都不可以提供完整的服务

说明：读写由 A、B 机房同时提供，A 机房到 B 机房网络断开后：

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 C 机房组成子区域，满足多数派
  - B 机房有 Leader：B 机房仍然可以和 C 机房组成子区域，满足多数派
- 从应用角度看：
  - A 机房无法访问 B 机房的数据，A 机房可能可以访问 A 机房的数据（取决于 PD Leader 所在）
  - B 机房无法访问 A 机房的数据，B 机房可能可以访问 B 机房的数据（取决于 PD Leader 所在）



因此，A、B 机房都不可以提供完整的服务，仅可能提供 Region Leader 位于自身时的数据服务



第二阶段：B 机房和 C 机房间网络断开（完全断开 B）。

预期情况：A 机房可以提供完整的服务，B 机房无法提供服务

说明：读写由 A、B机房同时提供，B机房被完全隔离

- 从 Region Leader 角度看
  - A 机房有 Leader：A 机房仍然可以和 C 机房组成子区域，满足多数派
  - B 机房没有有 Leader
- 从应用角度看：
  - A 机房可以访问 A 机房的数据
  - B 机房不可以访问 A 机房的数据

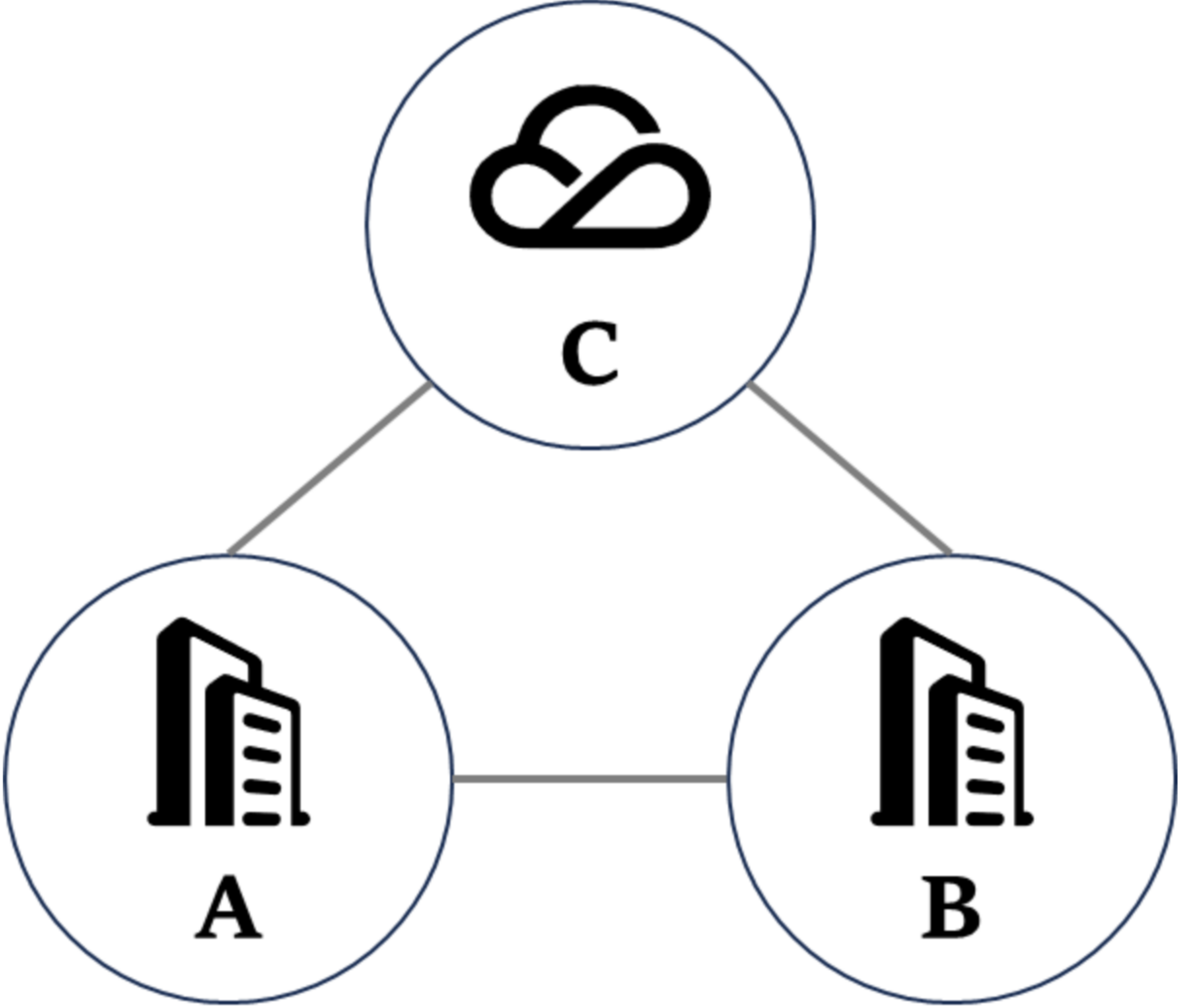
因此，A 机房可以提供完整的服务，B 机房无法提供服务

需要注意的是：B 机房历经从半断开到完全断开

- 在半断开时：A、B 机房都无法访问对方的数据，但可能能读写自己的数据（取决于 PD Leader 的位置）。此时写入 B 的数据在 B、C 上有相同的副本，A上没有。
- 当 B 完全断开时：B 不满足多数派，其上的 region leader 由 A、C 重新发起选举，但由于 A 上的部分数据不是最新，只能由 C 先成为 leader，在补完数据后再将 leader 转移给 A（机制上每60秒检查一次，进行 leader 的转移）
- 多中心架构中 C 机房按建议配置 raft-min-election-timeout-ticks/raft-max-election-timeout-ticks 参数限制 C 上的副本发起选举时间，即该极端场景下需等待 raft-max-election-timeout-ticks时间后，C 才会发起选举，并成为leader

因此，raft-min-election-timeout-ticks/raft-max-election-timeout-ticks 不建议设置的过大，建议设置在 60S 以内。

## 总结





部署拓扑	断开情况	集群状况	断开情况	集群状况
单区域多 AZ (对等三中心)	半断开 C (A≠C, C 可承载服务)	B 机房可提供完整的服务; A 机房无法提供完整的服务, 其访问 B 时正常, 而其访问 C 时会报错。	完全断开 C	A、B 机房均可提供完整的服务
双区域三中心 (带仲裁双中心)	半断开 B (A≠B)	A、B 机房都不可以提供完整的服务; 仅能提供本地数据服务 (要求 Region Leader 和 PD Leader 都位于自身机房)	完全断开 B	A 机房可以提供完整的服务 B 机房无法提供服务
	半断开仲裁 C (A≠C)	A、B 机房均可提供完整的服务	完全断开仲裁 C	A、B 机房均可提供完整的服务


版权声明：本文为 TiDB 社区用户原创文章，遵循 **CC BY-NC-SA 4.0** 版权协议，转载请附上原文出处链接和本声明。

### 评论

T

添加评论

评论



程序员小王 2024-09-25 14:50

半断开 和不半断开在业务有什么区别 不太理解

回复

<

1

>

#### 互助与交流

- 活动
- 问答论坛
- TiKV 社区
- Chaos Mesh 社区

#### 学习与应用

- 文档
- 专栏
- 视频课程
- 考试认证
- 典型案例
- 开发者指南

#### 发现社区

- TiDB User Group
- 问答之星
- 社区准则
- 联系我们
- 电子书

