BCM存储架构演进

简怀兵

Agenda

问题背景

1 → 2 → 3 → 4

问题背景

简要介绍BCM产品
初期的架构及存储系统
存在的瑕疵

解决方案

针对存在的不足
和BCM的中长期目标，
提出解决方案

技术挑战

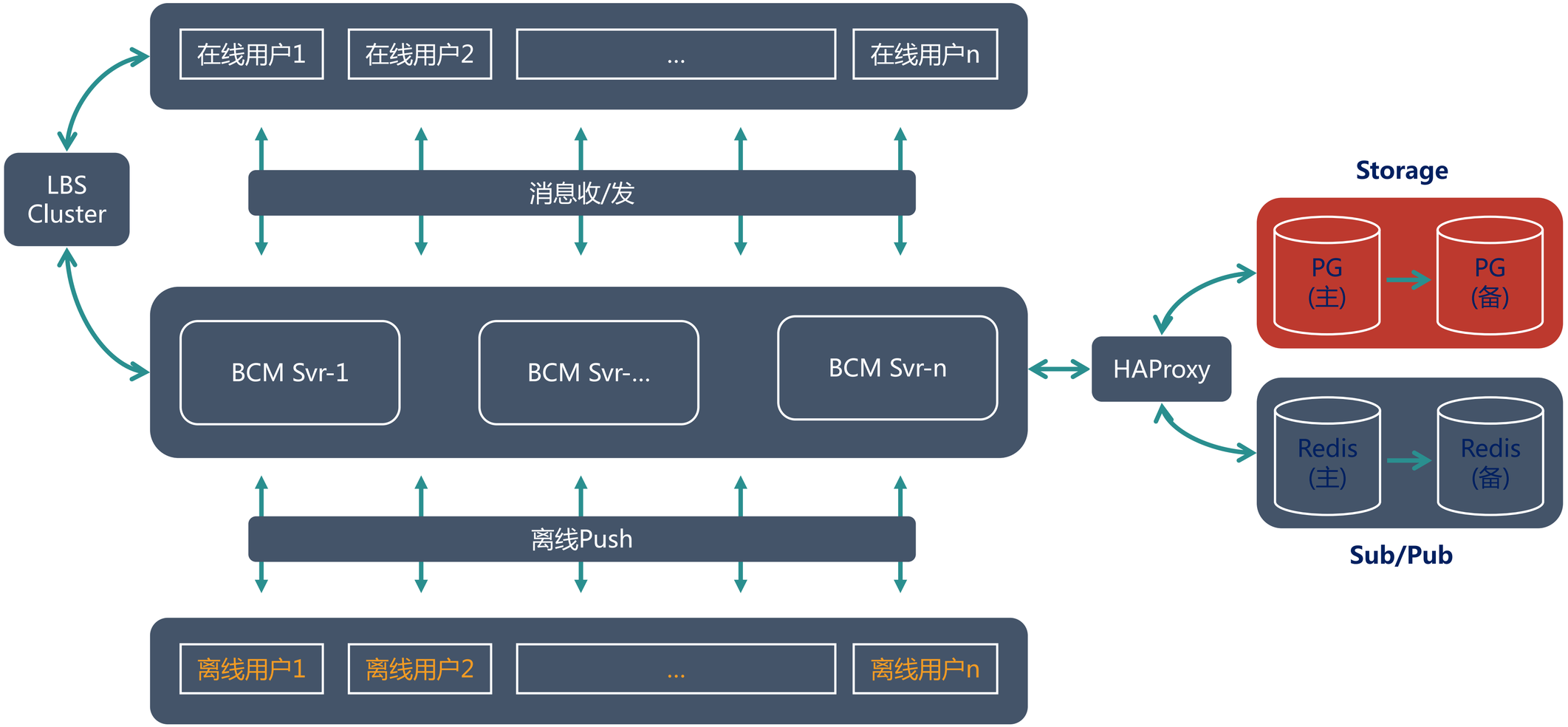
落地过程中面对的
技术挑战、业务诉求和
工程挑战及对策

未来展望

BCM存储体系的
下一步目标和计划

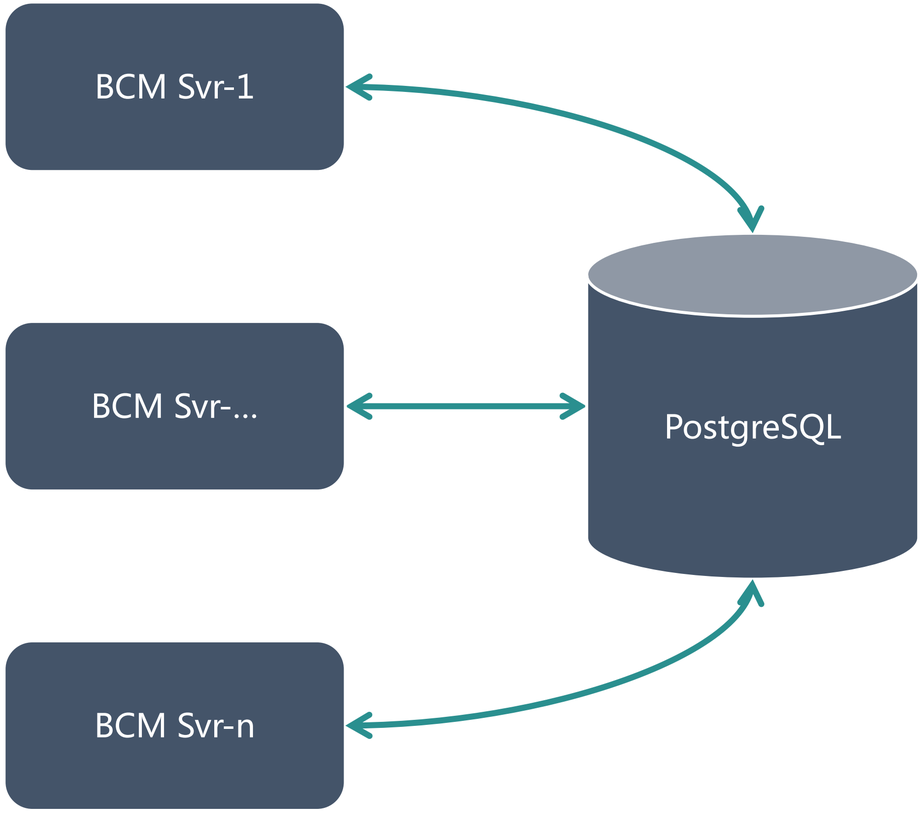
BCM业务架构V1

初期架构



BCM存储架构V1

集中式存储



诉求&痛点

不足 vs 迭代中的需求

扩展性不足

只能scale up
不能scale out
容错能力不足

业务特性

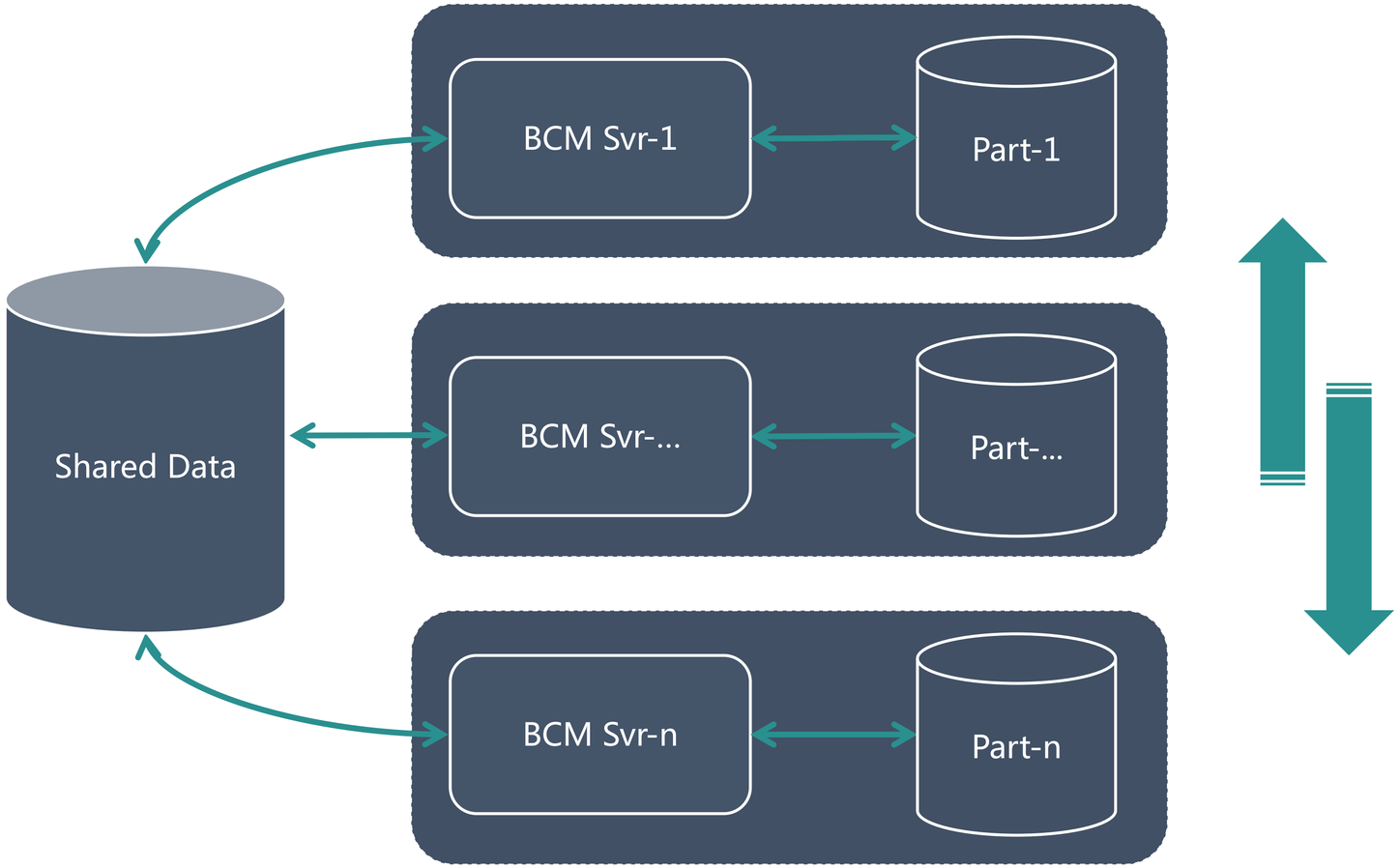
模型简单
实时性要求高、有TTL需求
跨region访问是常态
不允许dirty read

一致性模型单一

账户、登录信息需要强一致
消息只需要session级别一致

BCM存储架构V2

Scalability



Agenda

解决方案

1 → 2 → 3 → 4

问题背景

简要介绍BCM产品
初期的架构及存储系统
存在的瑕疵

解决方案

针对存在的不足
和BCM的中长期目标，
提出解决方案

技术挑战

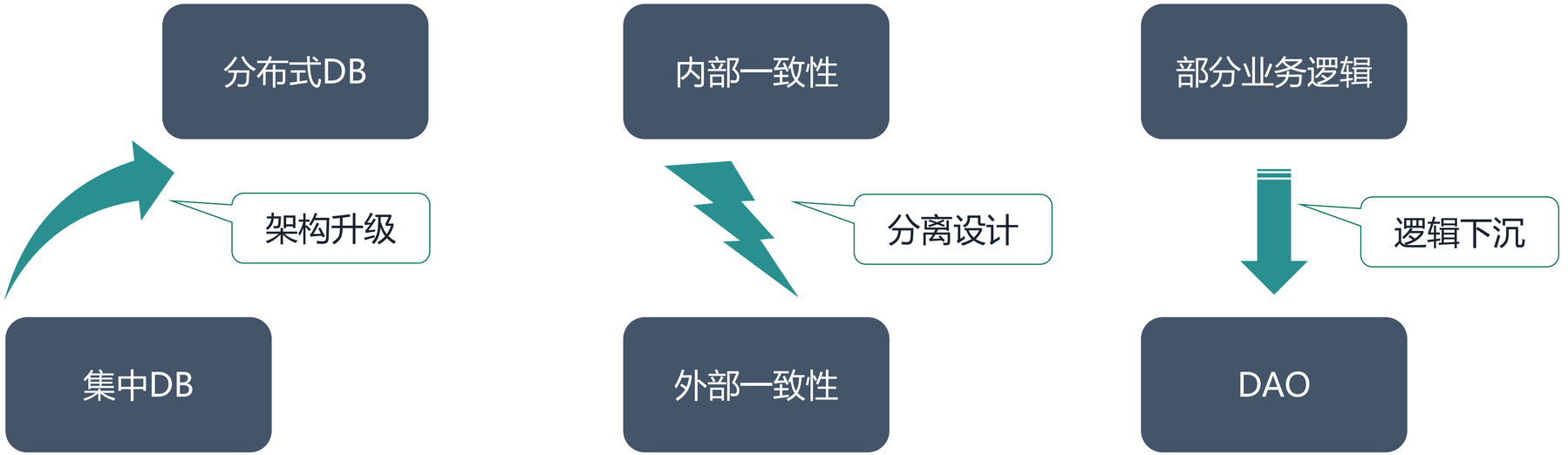
落地过程中面对的
技术挑战、业务诉求和
工程挑战及对策

未来展望

BCM存储体系的
下一步目标和计划

分析&思考

三种主要的设计手段



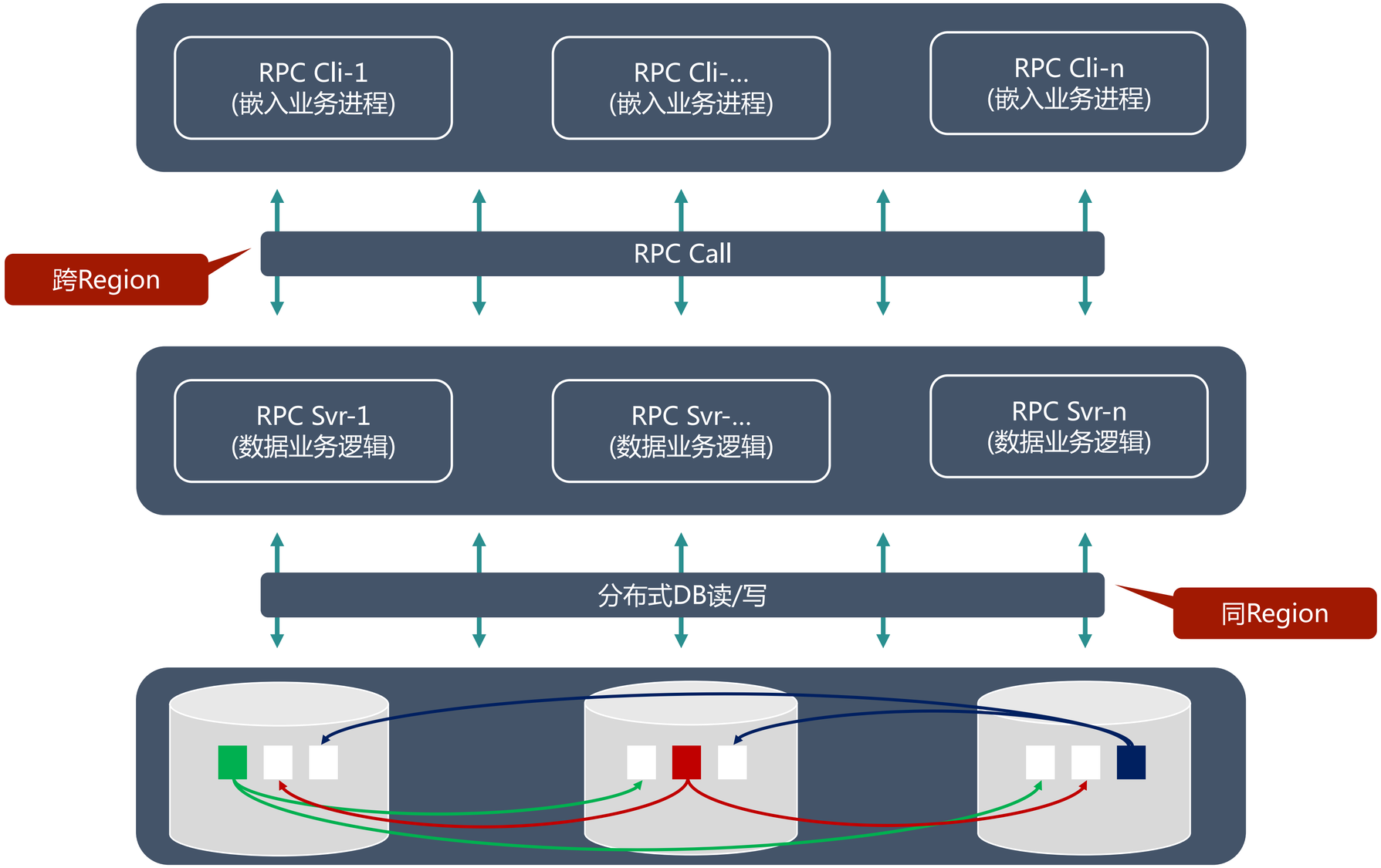
设计原则

PACELC

> If there is a partition (P), how does the system trade off availability and consistency (A and C); else (E), when the system is running normally in the absence of partitions, how does the system trade off latency (L) and consistency (C)?

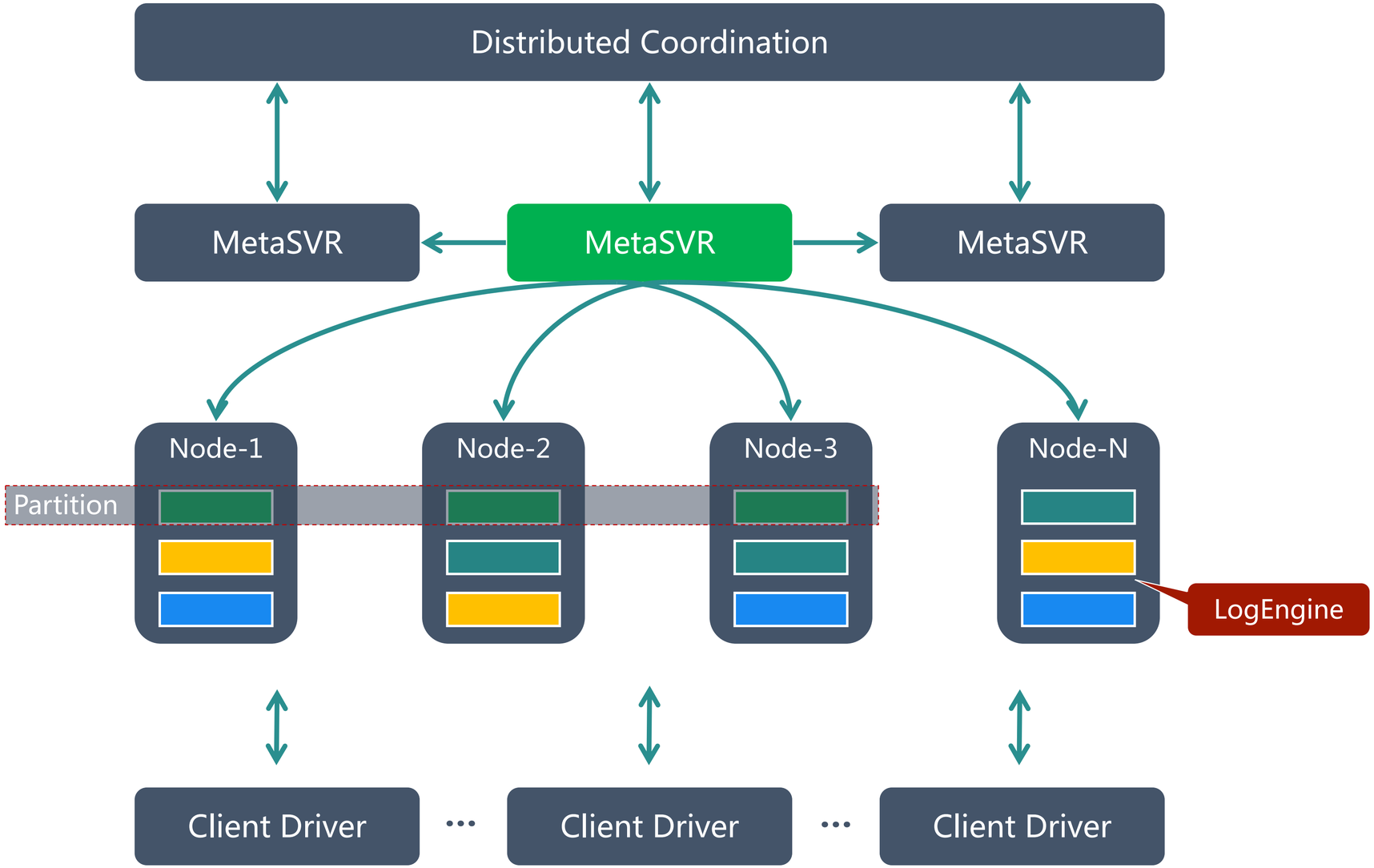
BCM layered存储架构

分层 & Stateless



BCM LogStore架构

A robust distributed key-value storage



Agenda

技术挑战

1 → 2 → 3 → 4

问题背景

简要介绍BCM产品
初期的架构及存储系统
存在的瑕疵

解决方案

针对存在的不足
和BCM的中长期目标，
提出解决方案

技术挑战

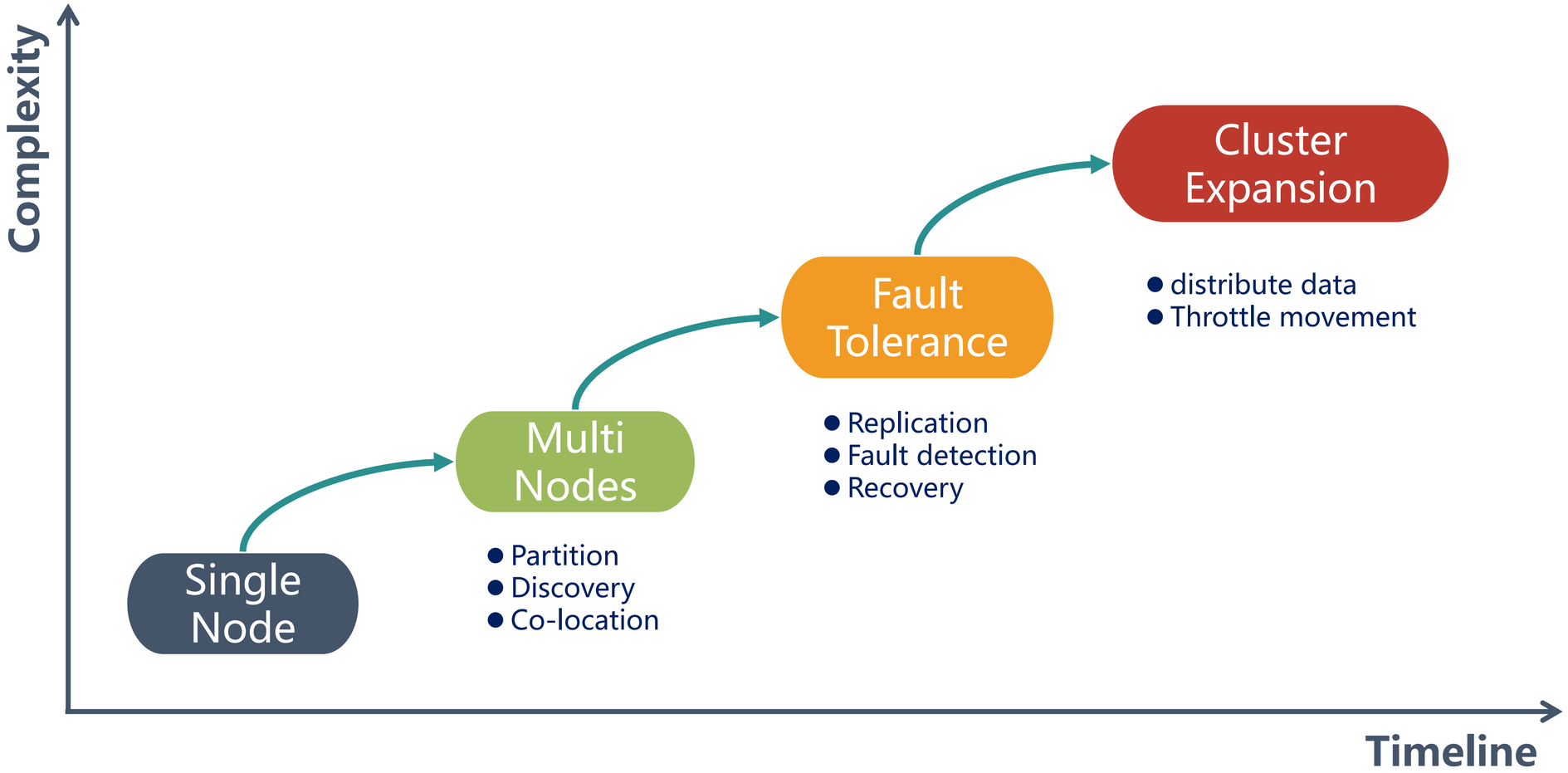
落地过程中面对的
技术挑战、业务诉求和
工程挑战及对策

未来展望

BCM存储体系的
下一步目标和计划

客观复杂度

系统能力 vs 复杂度



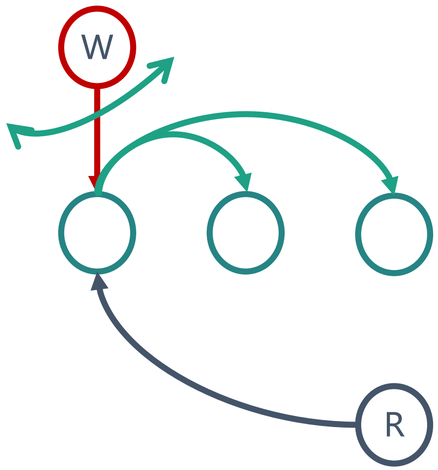
Replication Algorithm 1/2

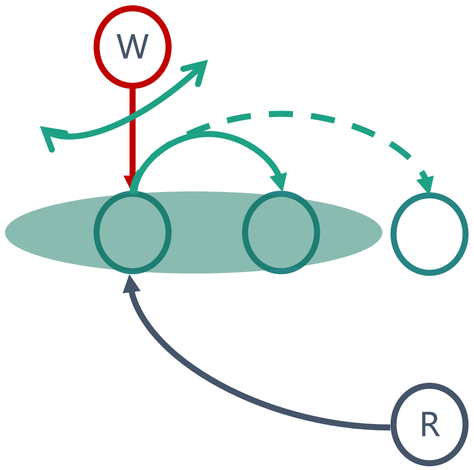
Flexible Replication

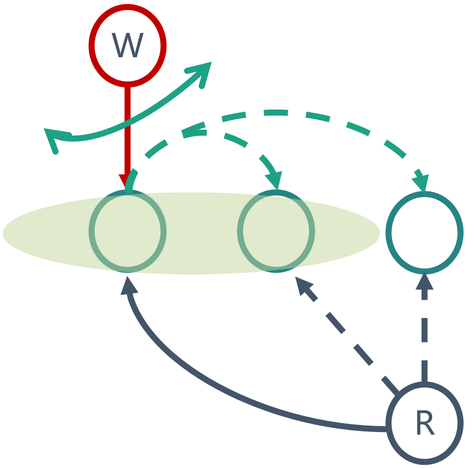
PacificA
ALL Ack

Paxos
Majority Ack

Flexible Paxos
Quorum Ack

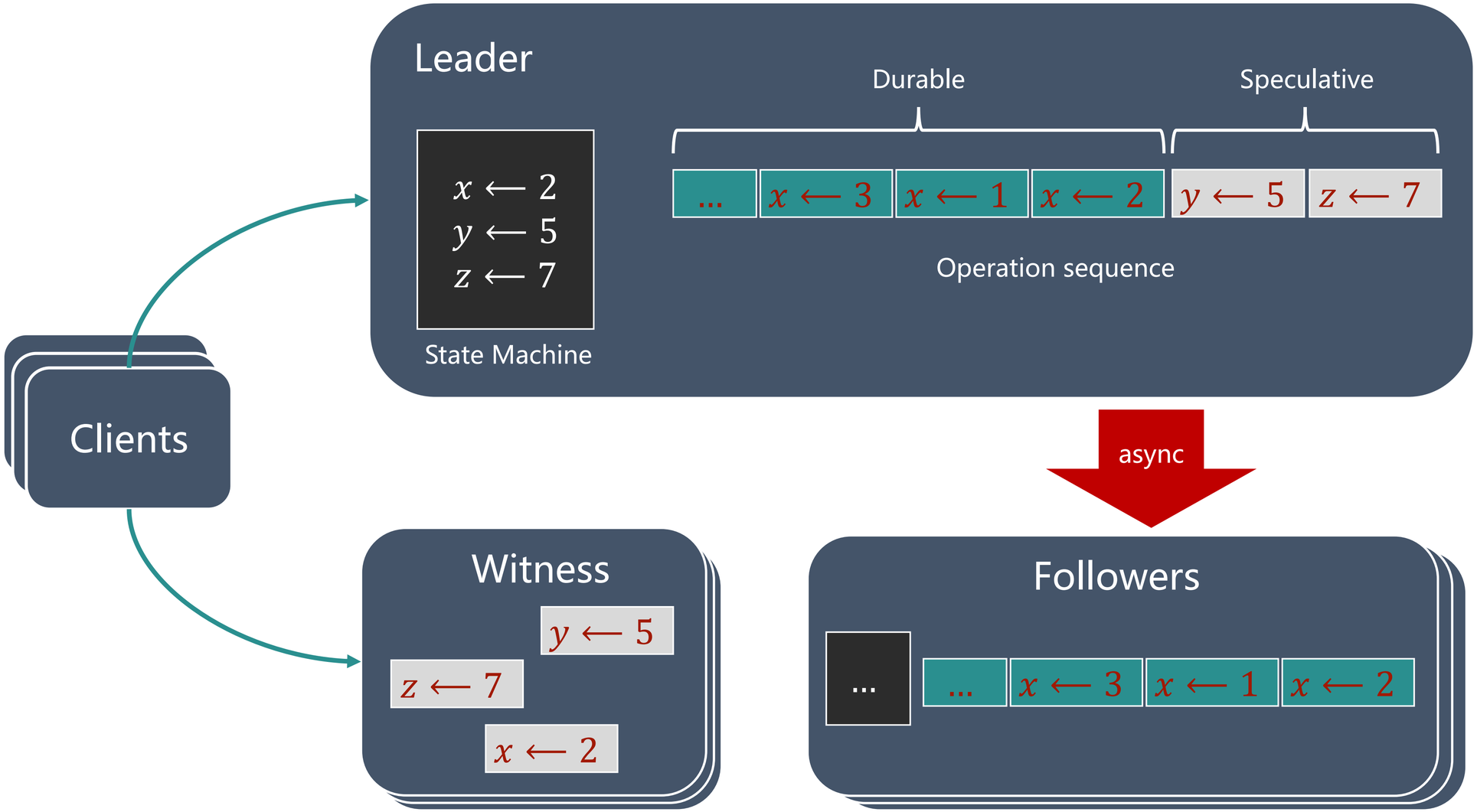






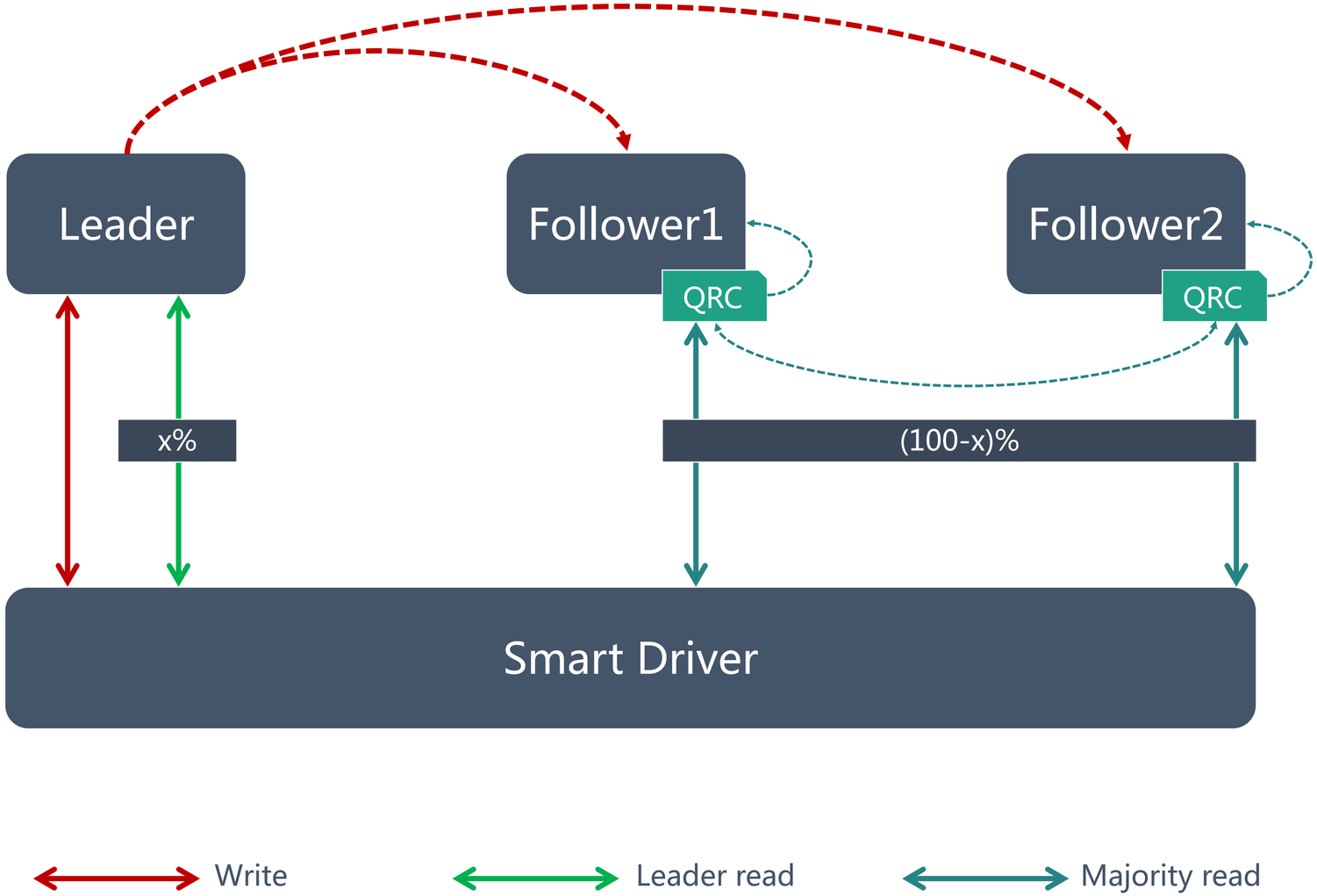
Replication Algorithm 2/2

Strong consistency with 1 RTT



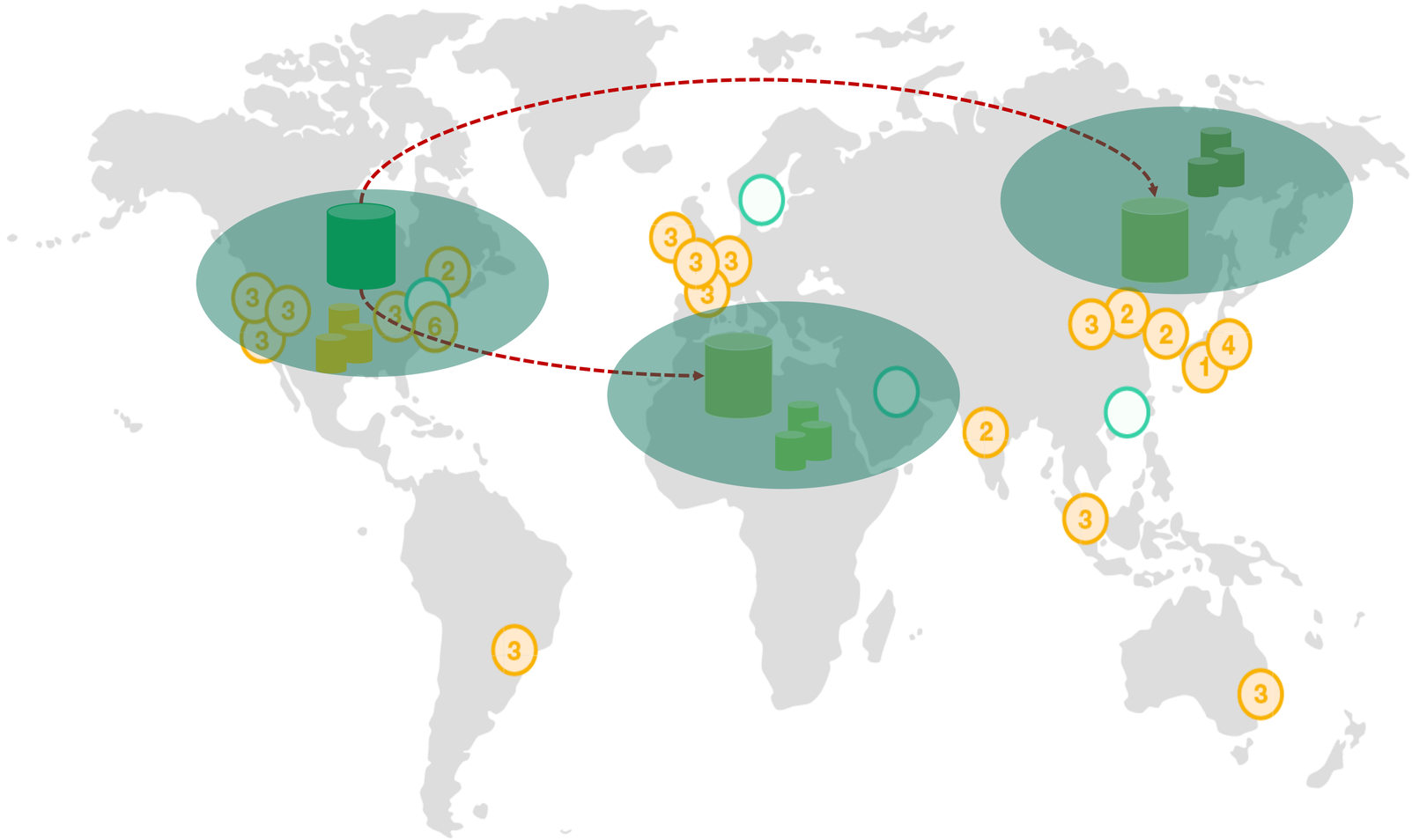
Hybird consistency model

Strong consistency and session consistency



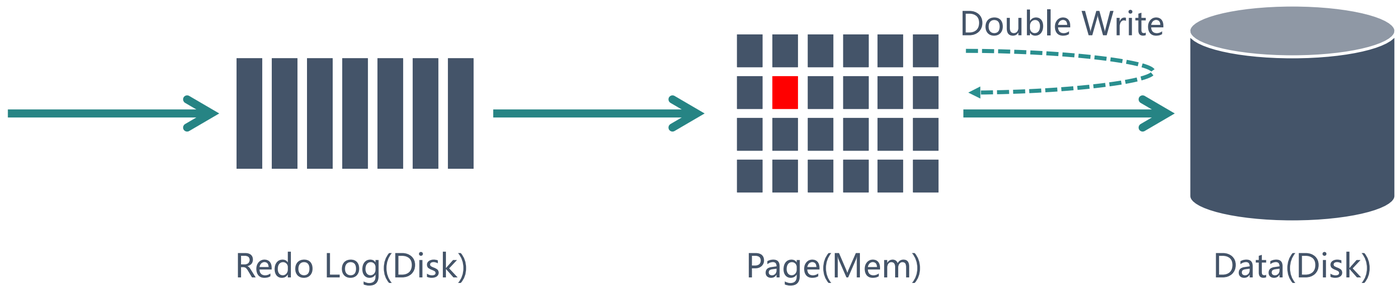
Location-aware Data Placement

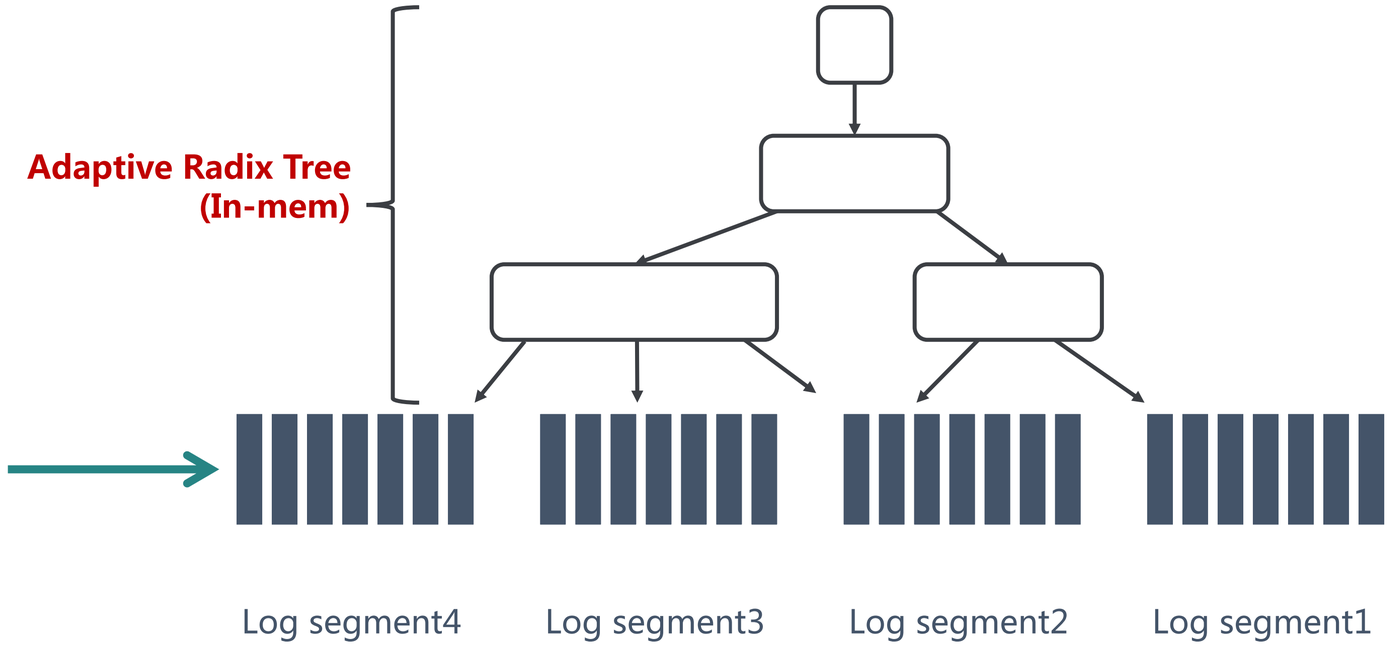
Follower read in local IDC



LogEngine: Write/read Amplification

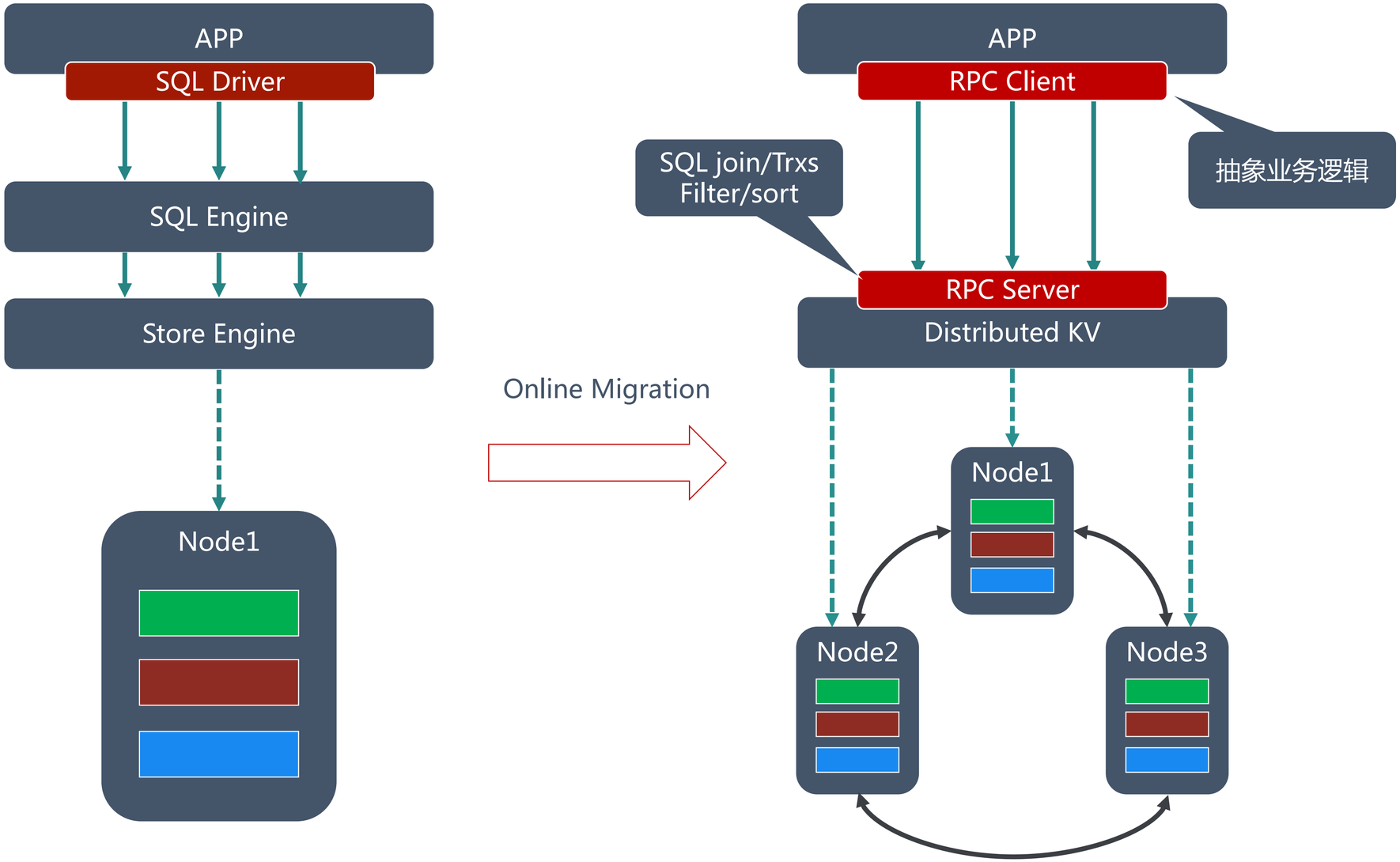
Log is data and ART index





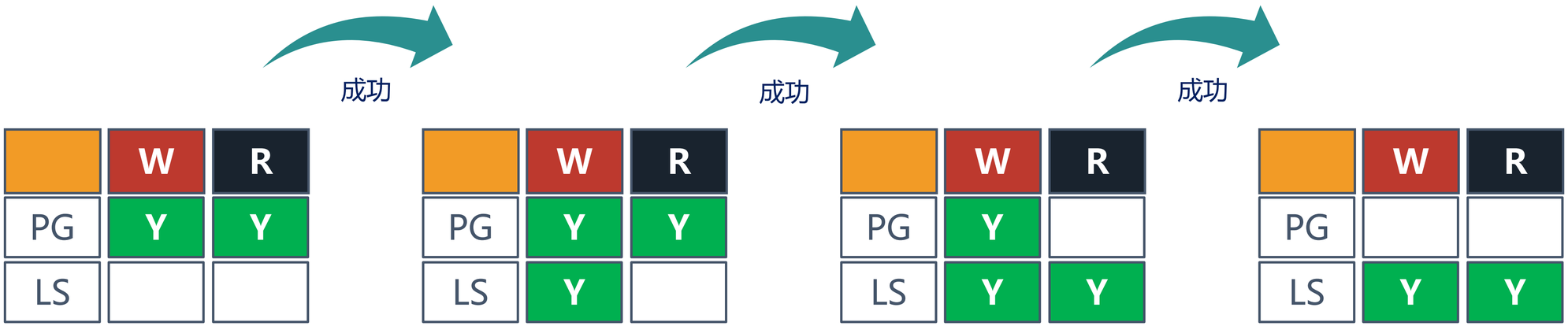
Online Migration 1/3

SQL to unified DAO



Online Migration 2/3

State Machine



PG: PostgreSQL

LS: BCM LogStore

Online Migration 3/3

Data storage strategy

id	username	email	deleted	name
1	alice	alice@example.com	0	alice
2	bob	bob@example.com	0	bob
3	bob2	bob2@example.com	0	bob2

key	value
1	username=alice,email=alice@example.com,deleted=0,name=alice
2	username=bob,email=bob@example.com,deleted=0,name=bob
3	username=bob2,email=bob2@example.com,deleted=0,name=bob

Store by tuple

key	value
username=alice,deleted=0	id=1
username=bob,deleted=0	id=2
username=bob2,deleted=0	id=3

复合unique index

key	value
name=alice,id=1	
name=bob,id=2	
name=bob,id=3	

non unique index

Agenda

未来展望

1 **问题背景**

简要介绍BCM产品
初期的架构及存储系统
存在的瑕疵

2 **解决方案**

针对存在的不足
和BCM的中长期目标，
提出解决方案

3 **技术挑战**

落地过程中面对的
技术挑战、业务诉求和
工程挑战及对策

4 **未来展望**

BCM存储体系的
下一步目标和计划

未来展望

Data storage strategy

LogStore加强

通用场景下的能力
pub/sub能力
Admin console

Open Source

LogStore作为一个分布式存储系统开源；LogEngine作为一个单机引擎开源

感谢聆听

—

THANKS!