大数据分布式系统并行设计与I/O优化

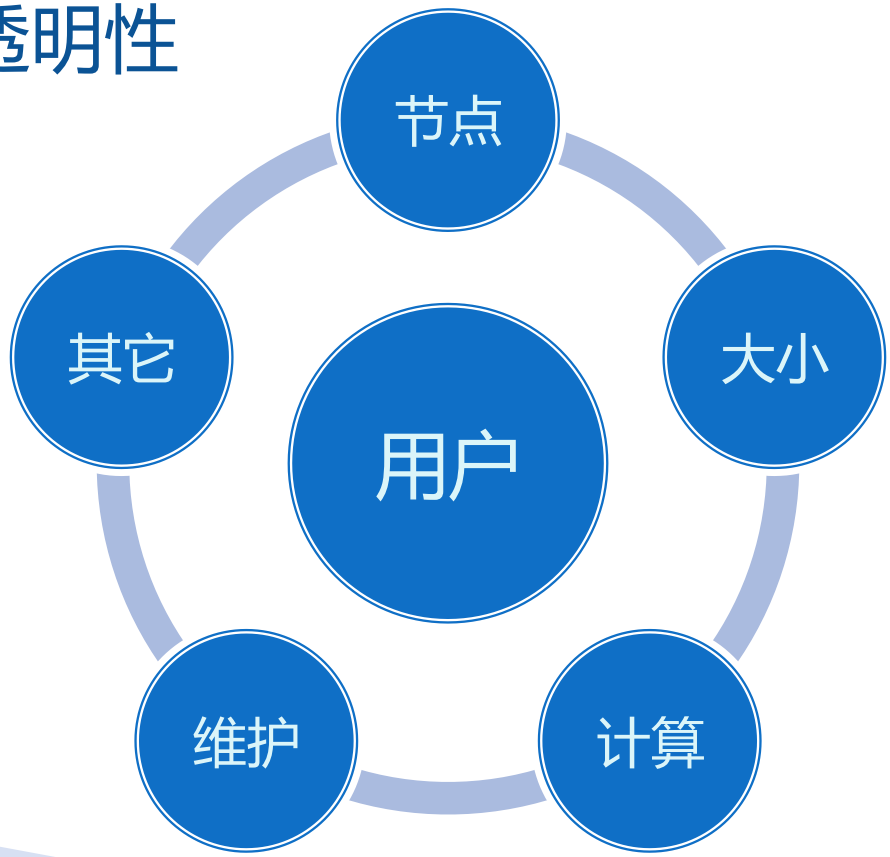
微博：@卢亿雷
邮箱：johnlya@163.com

提纲

- 分布式并行系统定义
- 分布式并行系统的特点
- 分布式并行系统生态图
- 分布式并行系统设计
- 案例分析
- 单台机器并行设计
- 分布式并行I/O优化
- 单台机器分布式化设想

分布式并行系统定义

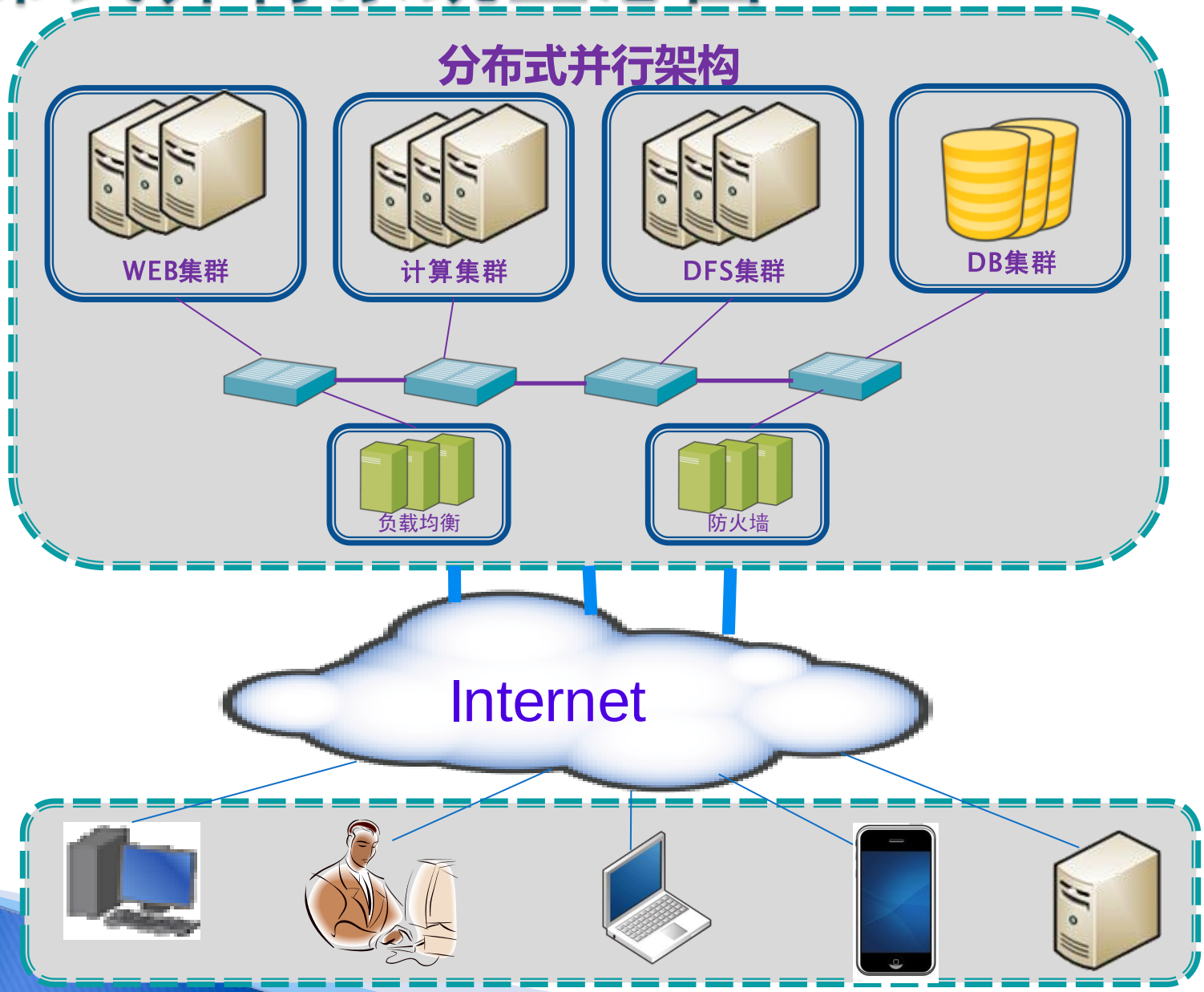
- 是多个系统的集合
- 子系统平行地相互作用
- 具有很好的低依赖性和透明性



分布式并行系统的特点

高可靠性	• 数据需要多份存储，保证数据不丢失
高可用性	• 提供7X24小时服务，保证服务不中断
高扩展性	• 提供透明升级扩容服务，保证服务不受限制
高性能	• 对高并发、低延迟有高要求，保证服务访问快速
高性价比	• 成本价格可控，尽量降低成本

分布式并行系统生态图



分布式并行系统生态图



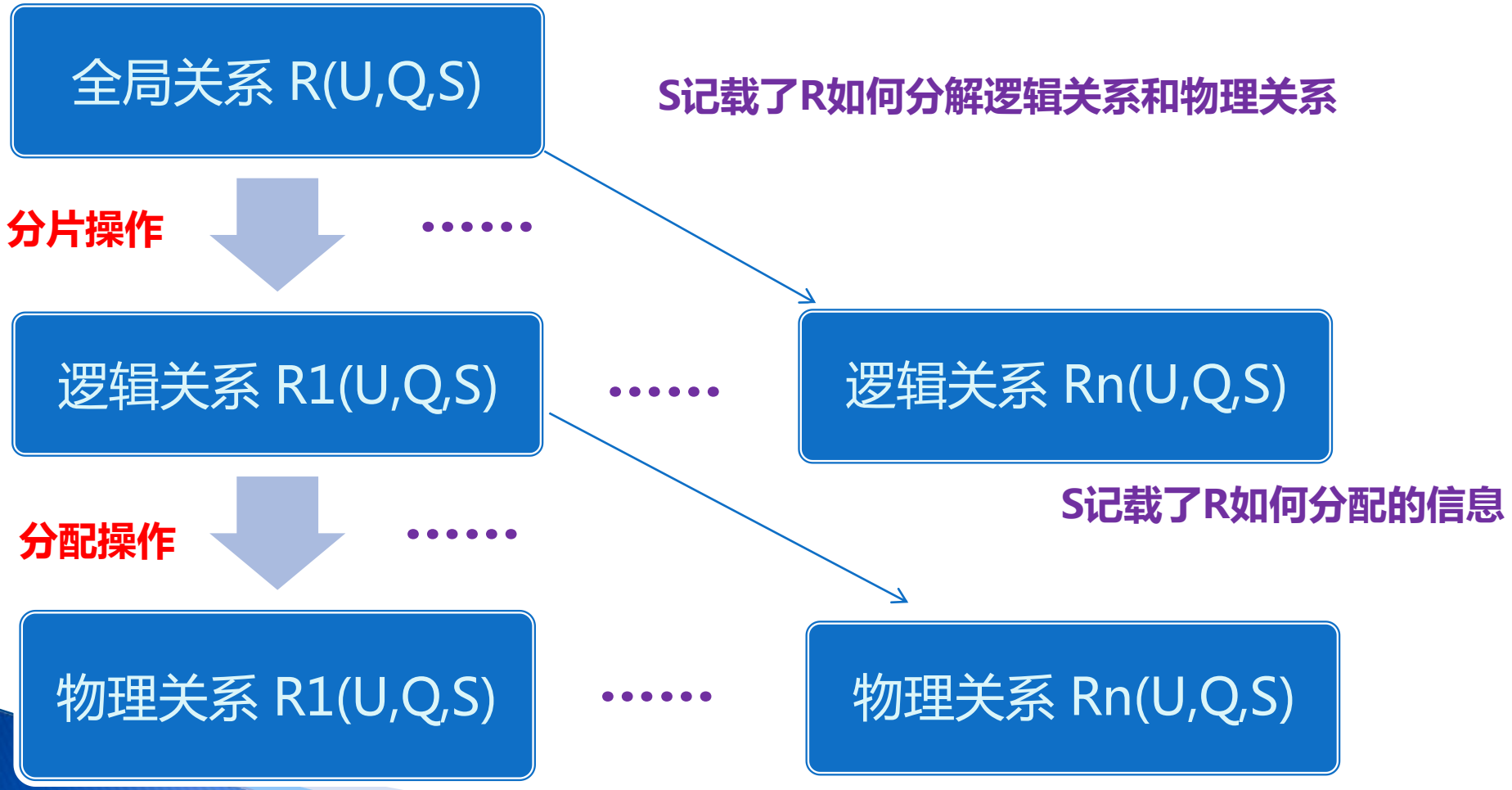

用户访问				
负载均衡	LVS	Nginx	HAProxy	.....
业务逻辑	权限管理	数据分发	数据挖掘	实时计算
网络协议	TCP		UDP	
操作系统	Unix	Linux	Windows	
硬件	CPU	内存	硬盘	网卡

分布式并行系统设计

- 全局（虚拟）关系
  - 由所有全局关系组成的子系统
- 逻辑关系
  - 由所有逻辑片段组成的子系统
- 物理关系
  - 由所有物理关系组成的子系统

分布式并行系统设计

- S是有关R被分片和分配的信息的集合



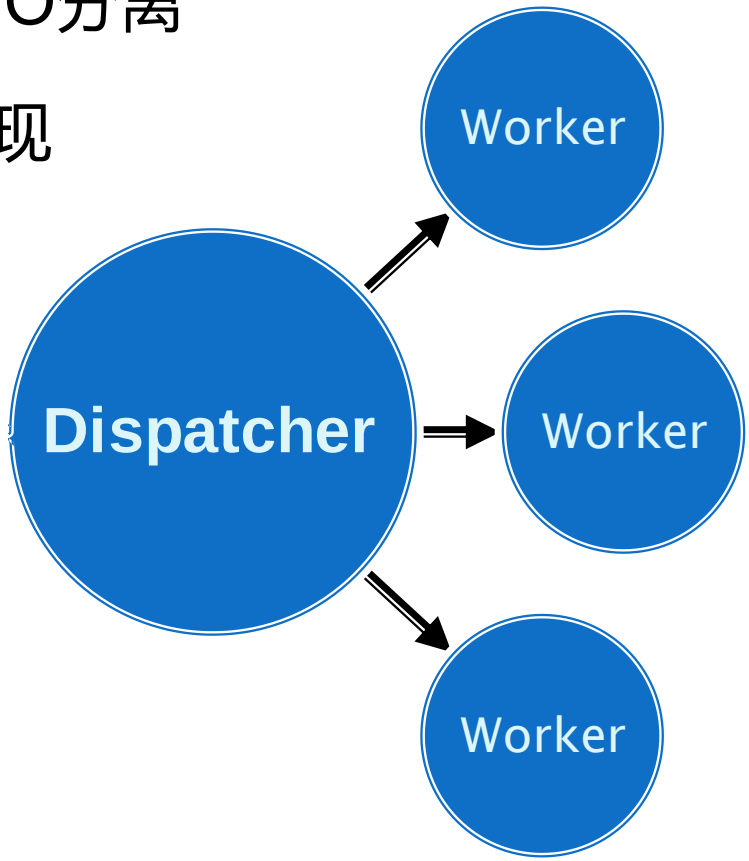
分布式并行系统设计模式

- 主从结构
  - 简单有效，结构清晰
  - 有单点失效问题
- 全对等结构
  - 无中心结点，无单点失效
  - 很难保证数据一致性

分布式并行系统设计模式

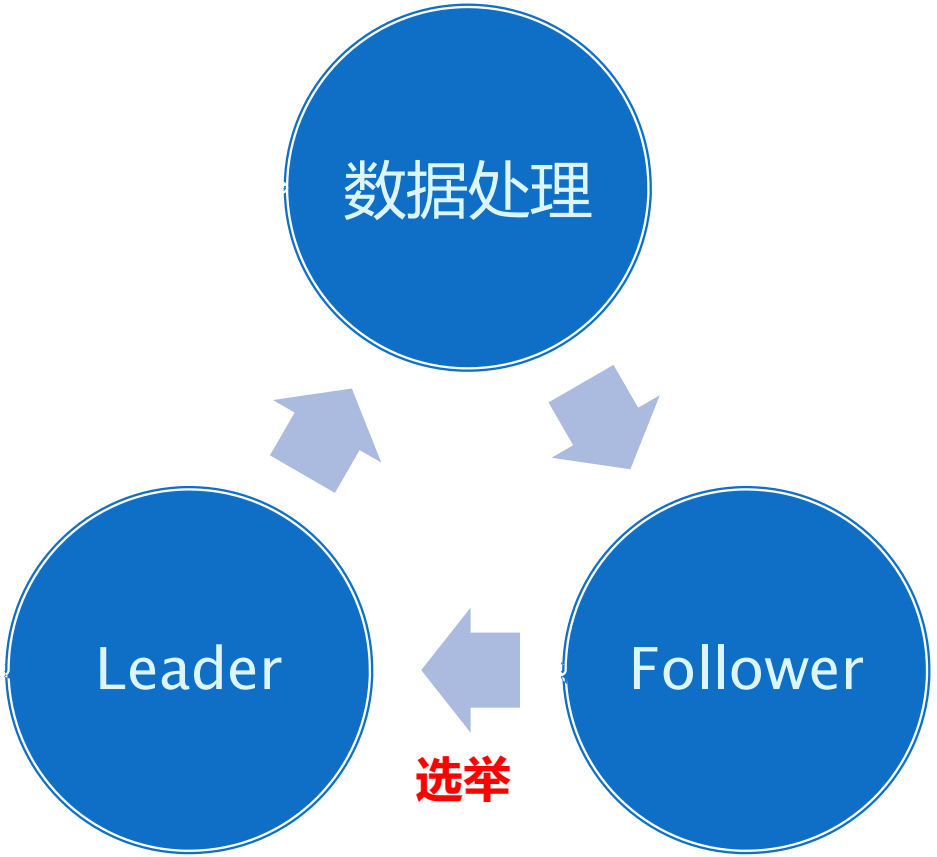
➢ 半同步/半异步（Half Sync/Half Async）

- 同步I/O和异步I/O分离
- 简化并发程序实现



分布式并行系统设计模式

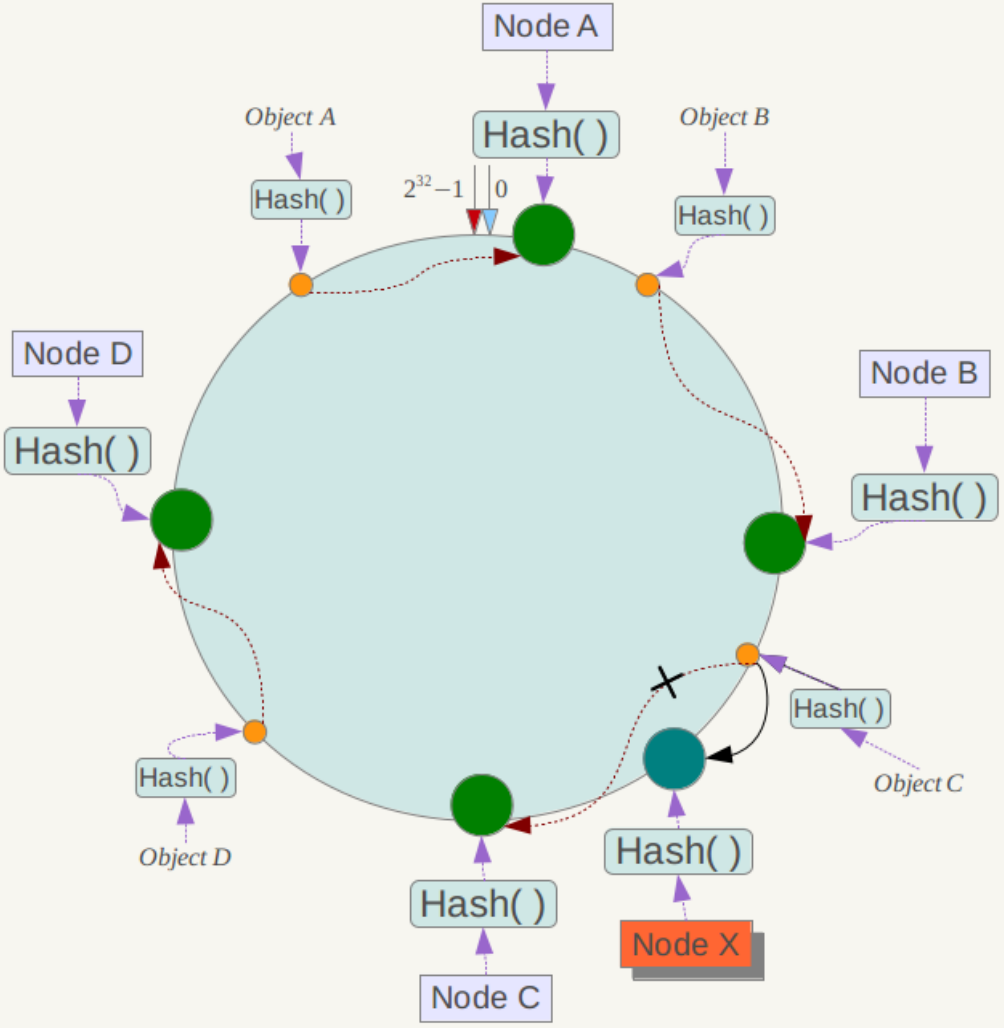
- 领导者/追随者模型（Leader/Followers）
  - 无需消息队列
  - 无上下文切换



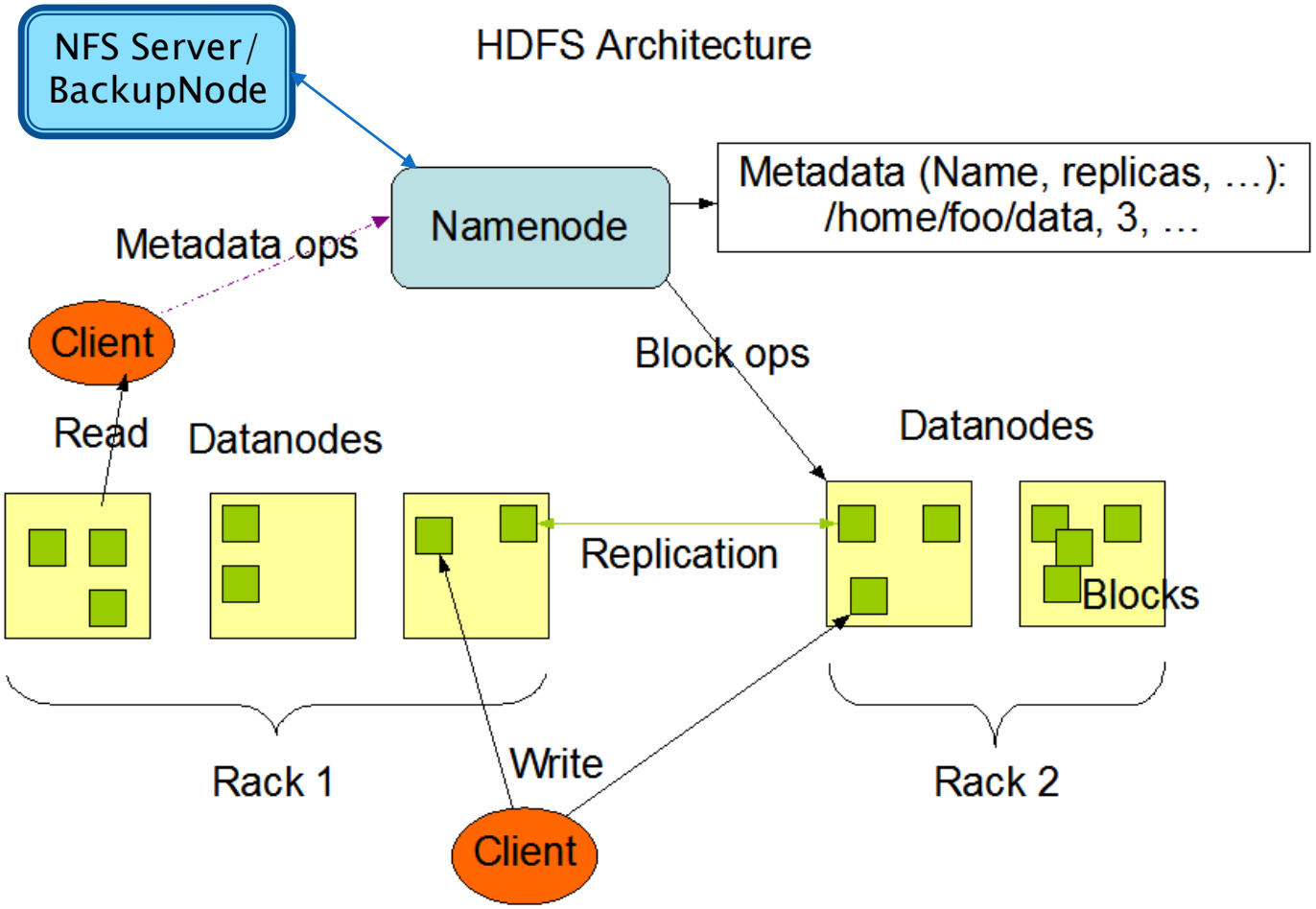
分布式并行系统设计模式

一致性哈希算法

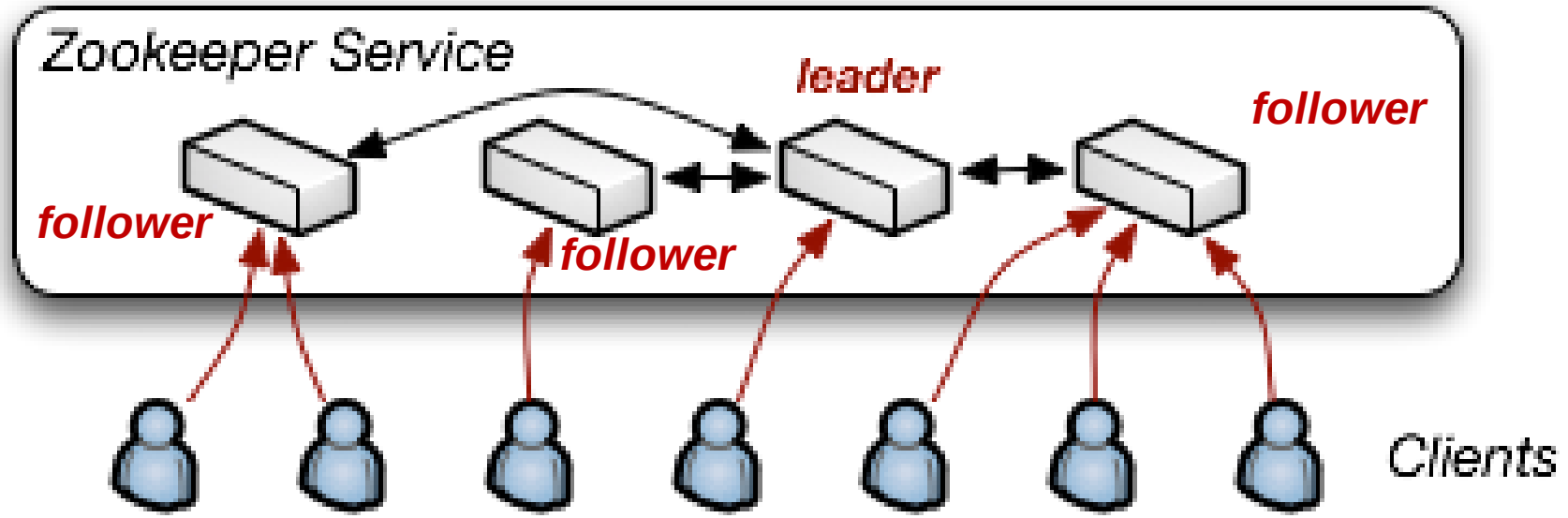
- 将整个哈希值空间组织成一个虚拟的圆环
- 只需重定位环空间中的一小部分数据



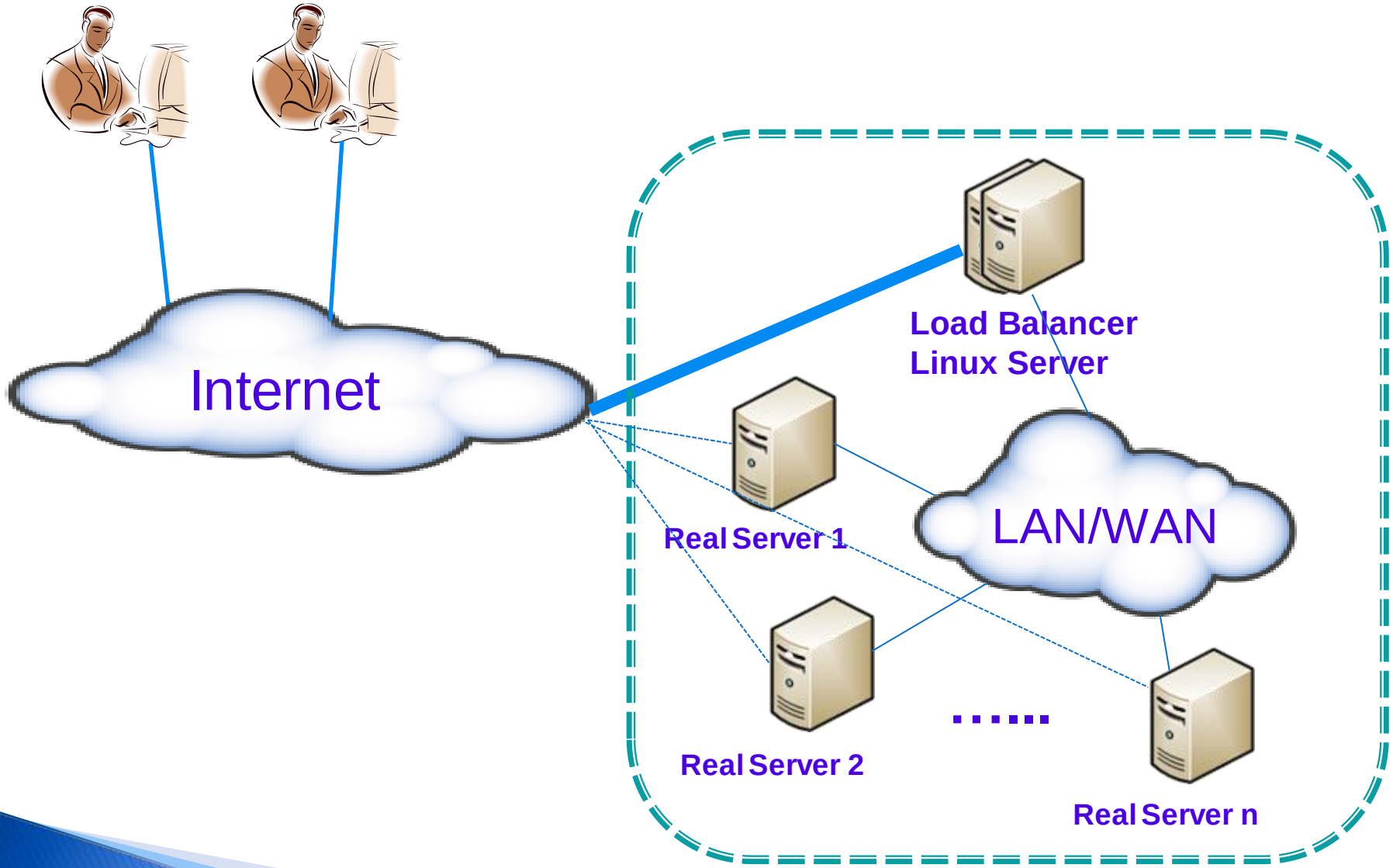
案例分析—分布式文件系统



案例分析—分布式服务框架



案例分析-LVS


Internet
Load Balancer
Linux Server
LAN/WAN
Real Server 1
Real Server 2
.....
Real Server n

单台机器并行设计

- CPU
  - 多进程多线程
  - Parallella：双核ARM A9加上64核多核浮点加速处理器达到90 GFLOPS
- 内存
  - 预先分配
  - 延时释放
  - 避免竞争

单台机器并行设计

- 磁盘
  - 多磁盘读写
- 网卡
  - 多网卡设置不同路由
- 文件系统
  - 设置文件系统缓存值
    - VFS索引节点缓存（Inode Cache）
    - 目录缓存（Inode Directory Cache）
    - 缓冲区缓存（Buffer Cache）
  - 设置文件系统预读值

分布式并行系统I/O优化

- 操作系统优化
  - 关掉不必要服务
  - 调整关键参数值，如连接数、文件系统格式化分块大小、文件系统预读及缓存值等
- 网络I/O策略优化
  - 区分带内与带外心跳
  - 控制备份速度
- 缓存策略优化
  - memcache
  - Redis
  - 缓存命中算法优化

分布式并行系统I/O优化

- 同步锁机制优化
  - 所有I/O操作的地方尽可能不要加同步锁
  - 大锁尽可能拆成小锁
- 多路复用I/O优化
  - 完成端口（IOCP）
  - Epoll模式
    - LT（level triggered）方式
    - ET （edge-triggered）方式

分布式并行系统I/O优化

- TCP选项优化
  - 发送实时指令
    - TCP_NODELAY
  - 缓存区大小设置
    - SO_SNDBUF
    - SO_RCVBUF
  - 延迟连接建立
    - TCP_DEFER_ACCEPT
- 内存操作优化
  - 避免内核空间和用户进程空间内存拷贝
    - sendfile
    - splice and tee
  - 预先分配，延时释放

单台机器分布式化设想

- 众核处理技术
  - 某个CPU坏了不影响系统
- 多内存条处理技术
  - 某个内存条或者某块内存不可用系统仍然可用
- 多硬盘处理技术
  - 自动过滤损坏硬盘或磁盘坏道
- 多网卡处理技术
  - 自动均衡网卡性能

Q & A

谢谢

分布式并行设计无处不在