促进软件开发领域知识与创新的传播

InfoQueue

QCon
全球软件开发大会

[上海] 2015年10月15-17日

ArchSummit
全 球 架 构 师 峰 会

[北京] 2015年12月18日-19日

关注InfoQ官方微信
及时获取ArchSummit演讲视频信息

ArchSummit全球架构师峰会深圳站2015

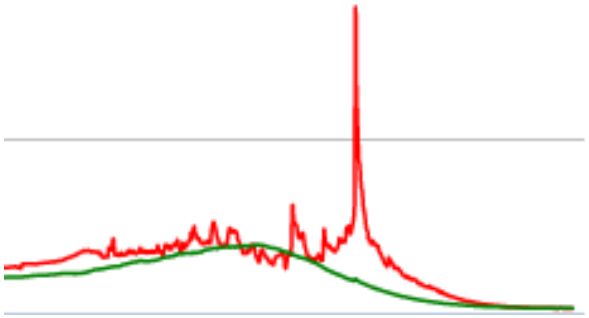
朋友圈技术之道

自我介绍

- 2012加入微信
- 负责
  - 帐号和关系链
  - 朋友圈
  - 消息

巨大的业务量

- 5.49亿微信MAU
- 朋友圈每天活跃量
  - 发表+赞+评论：过10亿
  - 浏览：过100亿
- 节日效应
  - 每个节假日都是一次狂欢：元旦、除夕、情人节、七夕、中秋、圣诞、 5/20、…
  - 节日流量：平时流量=2：1
- 突发效应
  - 零点：元旦、除夕、圣诞
  - 各类突发事件：重大体育比赛、北京下雪了、…
  - 突发峰值：平时峰值=5：1



敏捷的开发速度

- 每个季度一个主要版本

- 每天后台不定期变更

基础环境

- 全部采用C++
  - 正往C++11迁移
- 混合部署的普通服务器
  - 8/16/24/32 cores w/ hyper-thread
  - 8/16/32G memory
  - Hard disk/SSD
- 海量带宽

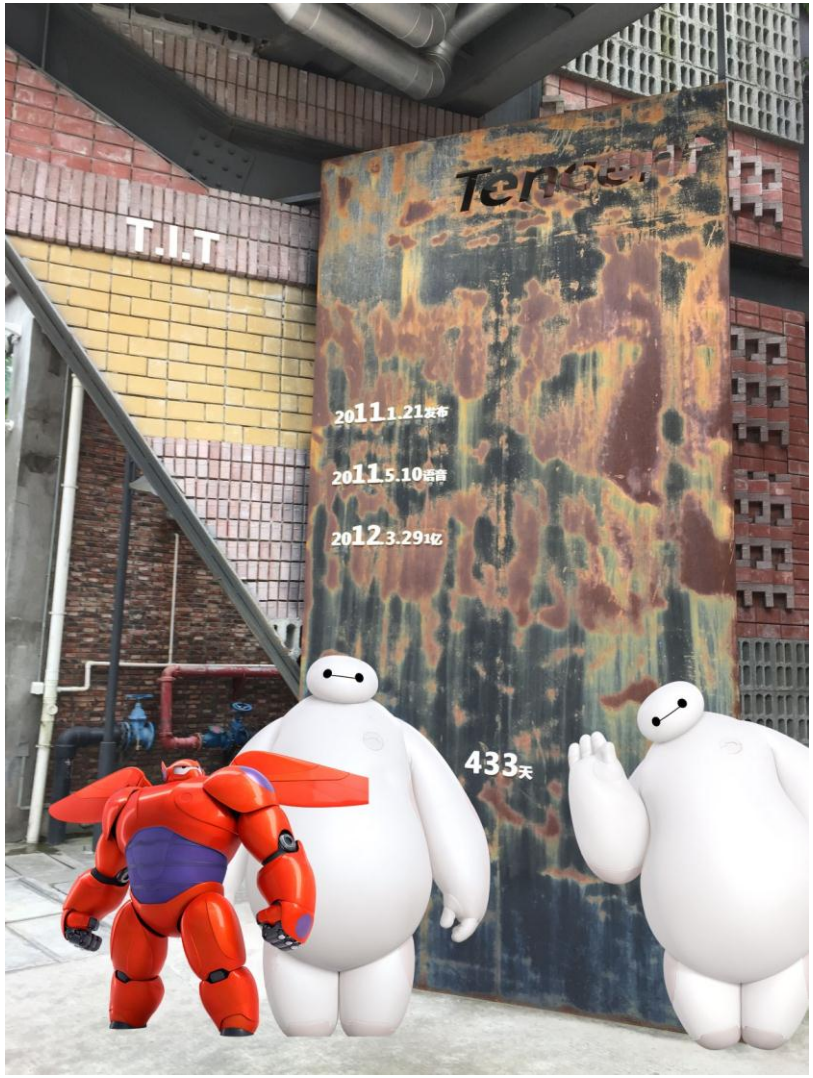
“？”团队

现在流行小团队

- Instagram
- What’s app
- …

“微”团队

- 人员增长了50%
- 从2位增长到3位



朋友圈“微”团队诚邀第“4”人加入

- 如果你想踏上移动互联网的巅峰
- 如果你想服务全球亿万用户
- 如果你想挑战世界级的分布式系统技术难题
- 如果你想和一群志同道合的码农一起愉快玩耍

请加入我们！

强大的基础设施：站在巨人的肩膀上

- 腾讯CDN
  - 图片和视频上载、存储、分发
- RPC 框架
- Key-value (KV)存储系统
- 强大、方便、灵活的部署系统

强大的RPC框架

- C++框架
- 支持protobuf描述接口
- 支持进程/线程/协程多种模式
  - 支持数十万的并发协程
  - 方便编写和调试“同步”的网络调用和服务
- 从CGI到叶子服务器，全系统透明支持过载保护和QoS
- 可为每次CGI调用自动生成全系统的调用关系图call graph
- 每个服务自动内建500多个监控项
  - 各个接口的客户端和服务器调用量
  - 接口耗时分布
  - 接口返回码分布
  - 当前QoS服务状况
  - …

高性能分布式key-value存储系统

- 三机一组
  - 三机间支持数据强一致性
  - 容忍一台机出错、自动切换
- 三机分布在一个数据中心的三个独立园区
  - 在任一园区提供本区读写服务
  - 容忍一个园区网络隔离
- 详情参见去年相关讲座

微信架构

上海

深圳/香港/加拿大

统计/监控

发布/配置

接入层

逻辑层

存储代理层

存储层

接入服务器

注册/登录

消息

朋友圈

LBS

群聊

通知

批处理

好友推荐

账号

关系链

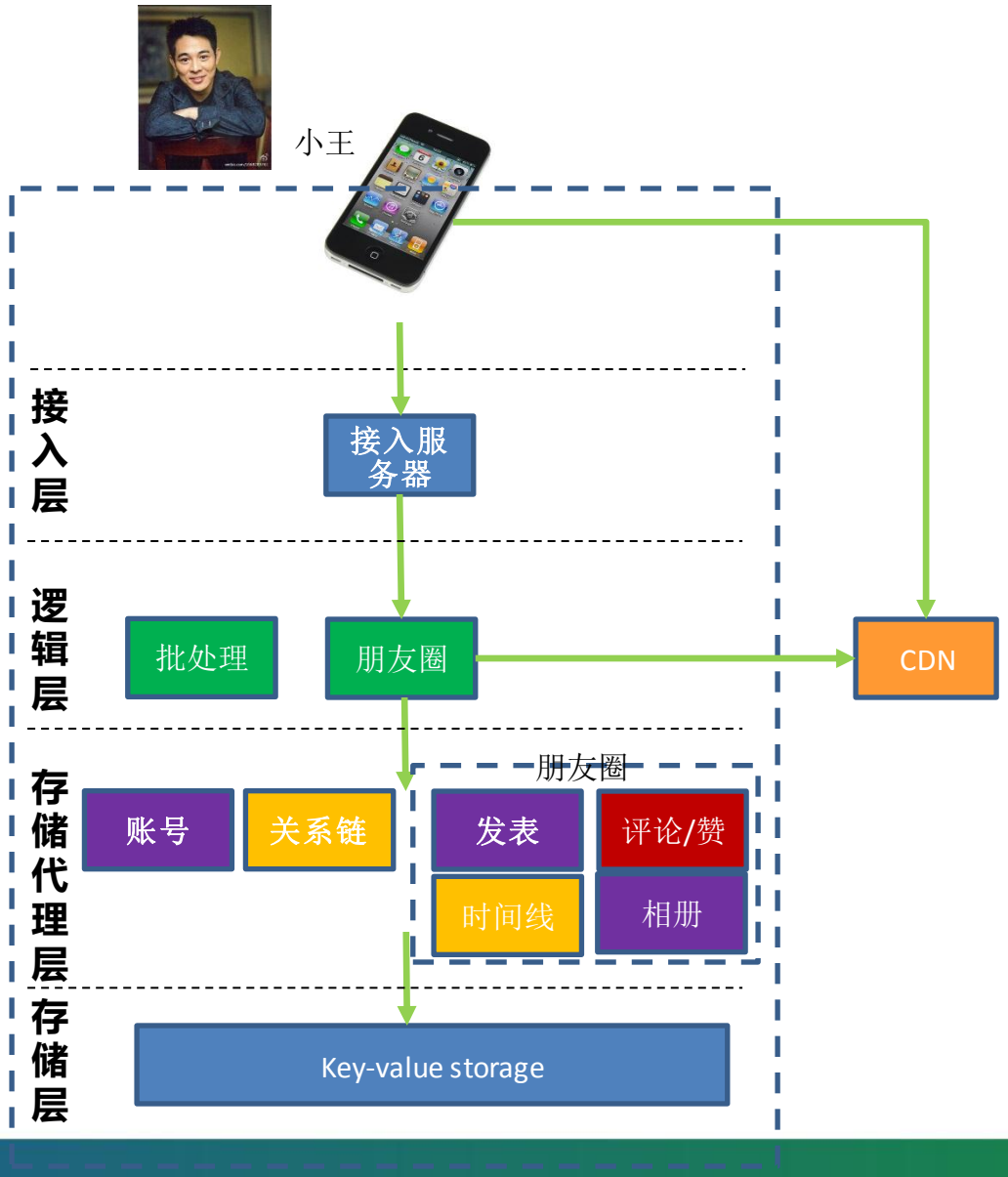
消息

群管理

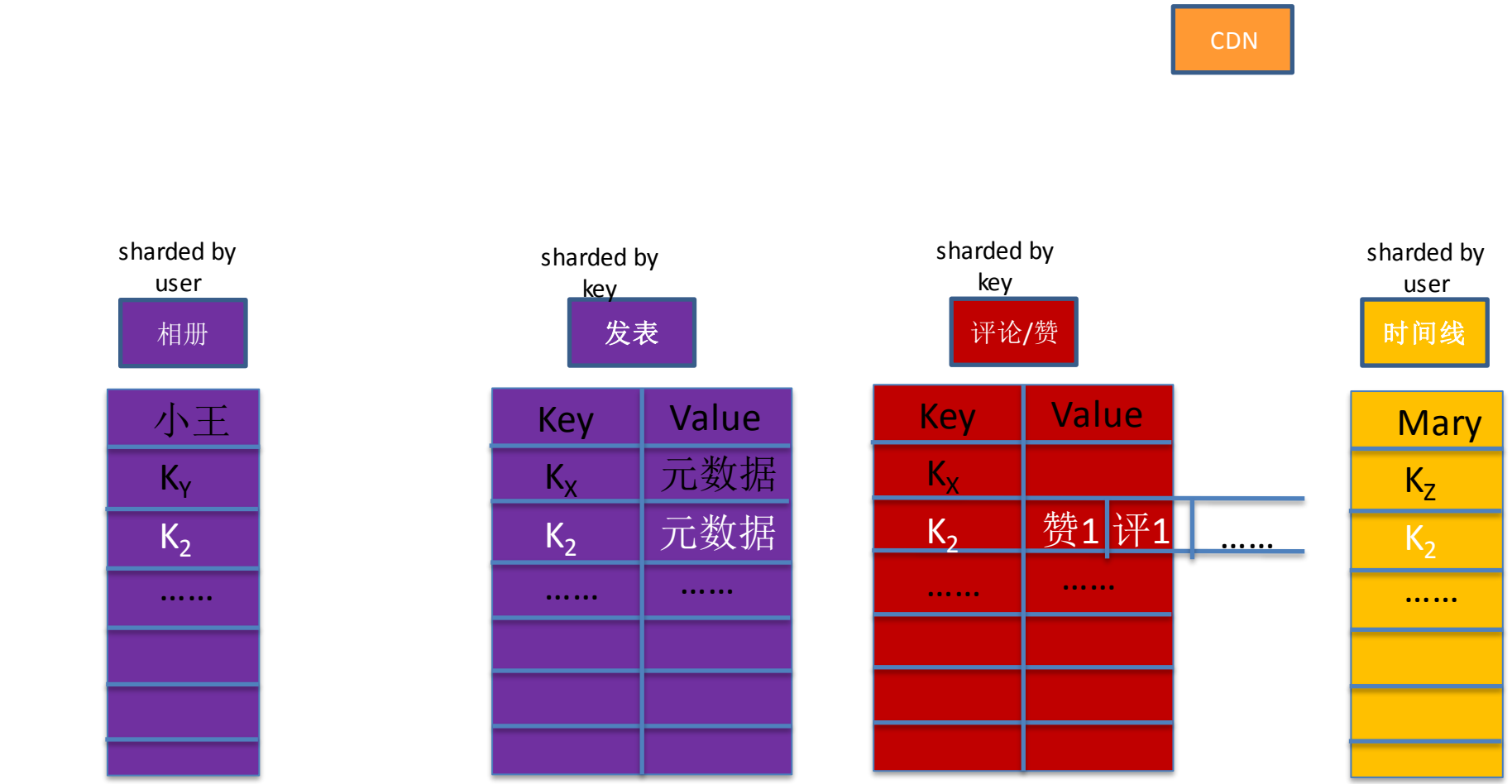
朋友圈

Key-value storage

朋友圈架构


小王
接入层
接入服务器
逻辑层
批处理
朋友圈
CDN
存储代理层
朋友圈
账号
关系链
发表
评论/赞
时间线
相册
存储层
Key-value storage

性能水平扩展


CDN
sharded by user
相册
小王
K_Y
K_2
……
sharded by key
发表
Key
Value
K_X
元数据
K_2
元数据
……
……
sharded by key
评论/赞
Key
Value
K_X
K_2
赞1
评1
……
……
……
sharded by user
时间线
Mary
K_Z
K_2
……

发表与浏览：数据单副本、索引写扩散/检查更新单读取

小王

Mary

1：上载

CDN

D：拉取图片

6

2

朋友圈 CGI

5：添加时间线更新任务

批处理

朋友圈 CGI

A

sharded by user

相册

3：添加新发表

sharded by key

发表

.....

5：给好友时间线添加新索引

sharded by user

时间线

B：拉取时间线更新

小王
K_Y
K_2
.....

4：添加新发表索引

Key	Value
K_X	元数据
K_2	元数据
.....	.....

C：拉取元数据

Mary
K_Z
K_2
.....

数据单副本、索引写扩散/检查更新单读取

- 数据（发表）单副本：减少存储开销
- 索引写扩散/检查更新单读取：减少检查更新时的读扩散

赞/评论、与浏览

小王

1：发表K_2

朋友圈 CGI

发表

Key	Value
K_X	元数据
K_2	元数据
……	……

Mary

2、3…：赞、评论

朋友圈 CGI

评论/赞

Key	Value
K_X	
K_2	赞1 评1 ……
……	……

……

……

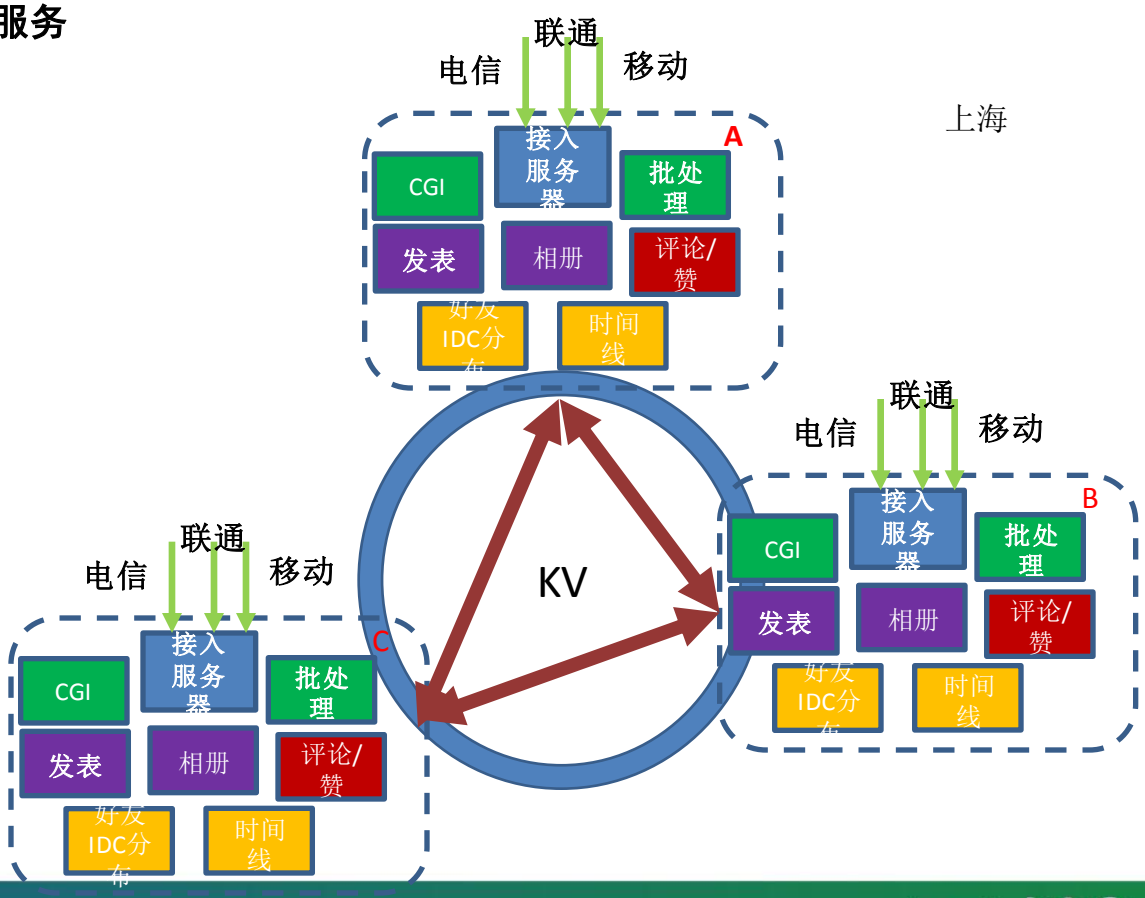
Kate

拉取评论

朋友圈 CGI

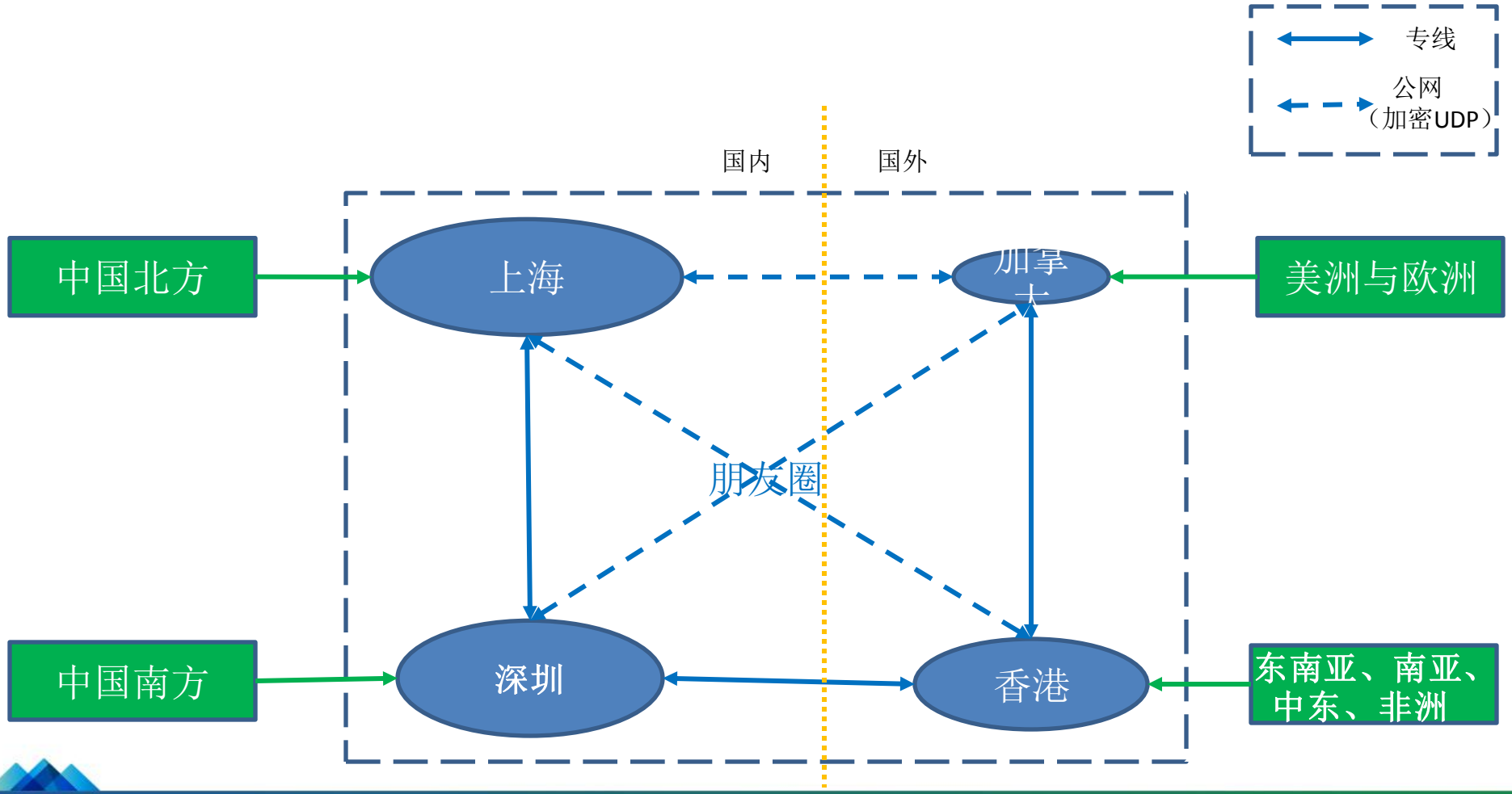
微信/朋友圈部署、接入、与容灾

- 1地3园区对等部署
  - 对等接入：每个园区同时接入电信、联通、移动等运营商
  - 数据对等分布与同步：KV提供园区数据对等部署与同步
  - 对等服务：每个园区服务能力均等
- 容灾
  - 任何两个园区都可提供全量无损服务
- 优点
  - 提供完全无损的容灾体验
  - 对逻辑层完全透明



从一地到全球：微信/朋友圈数据中心全球分布

- 每个区域独立服务本区域微信/朋友圈用户



朋友圈多区域自治架构

小王

Mary

颜色说明

全局数据

全局并发修改数据

本地用户数据

上海

香港

接入层

逻辑层

存储代理层

存储层

接入服务器

朋友圈

CDN

Idc queue

账号

关系链

发表

评论/赞

时间线

相册

Key-value storage

核心数据读写分析

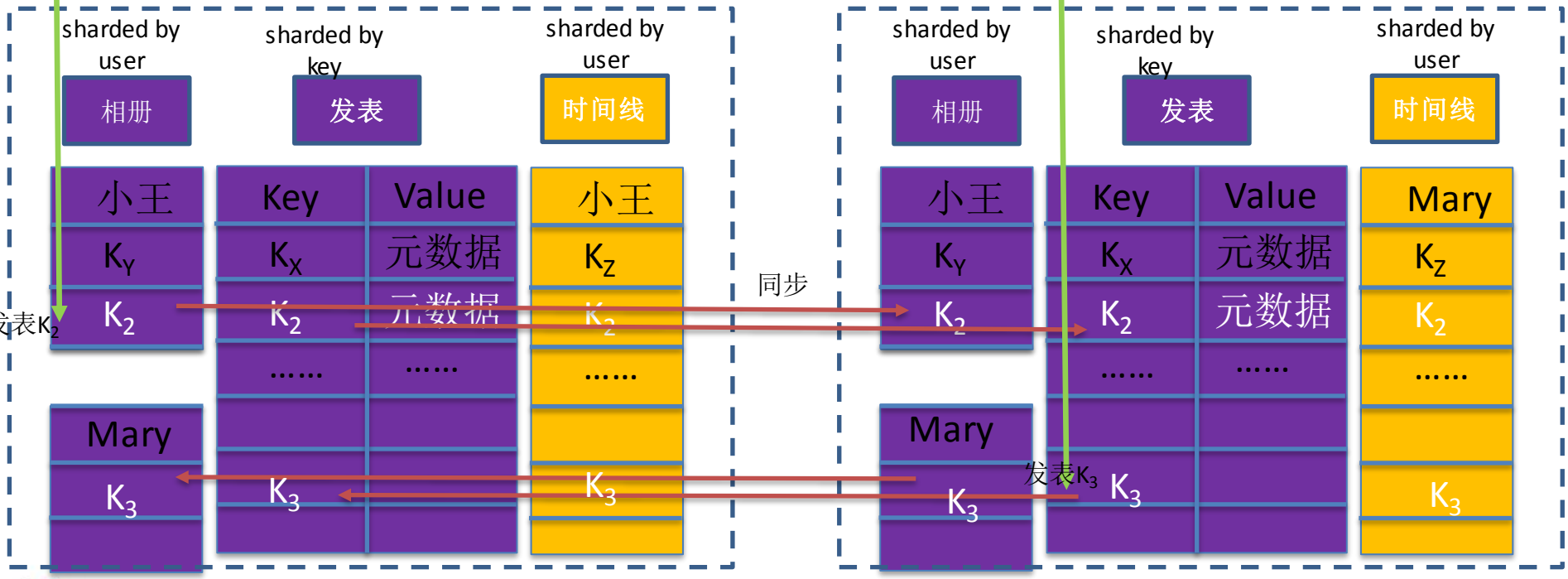
- 相册与发表
  - 本地idc写入、单向同步到其它idc
  - Key的全局唯一性保证无冲突
- 时间线
  - 本地idc只保存本地用户时间线的key，无需同步

小王

上海

Mary

香港

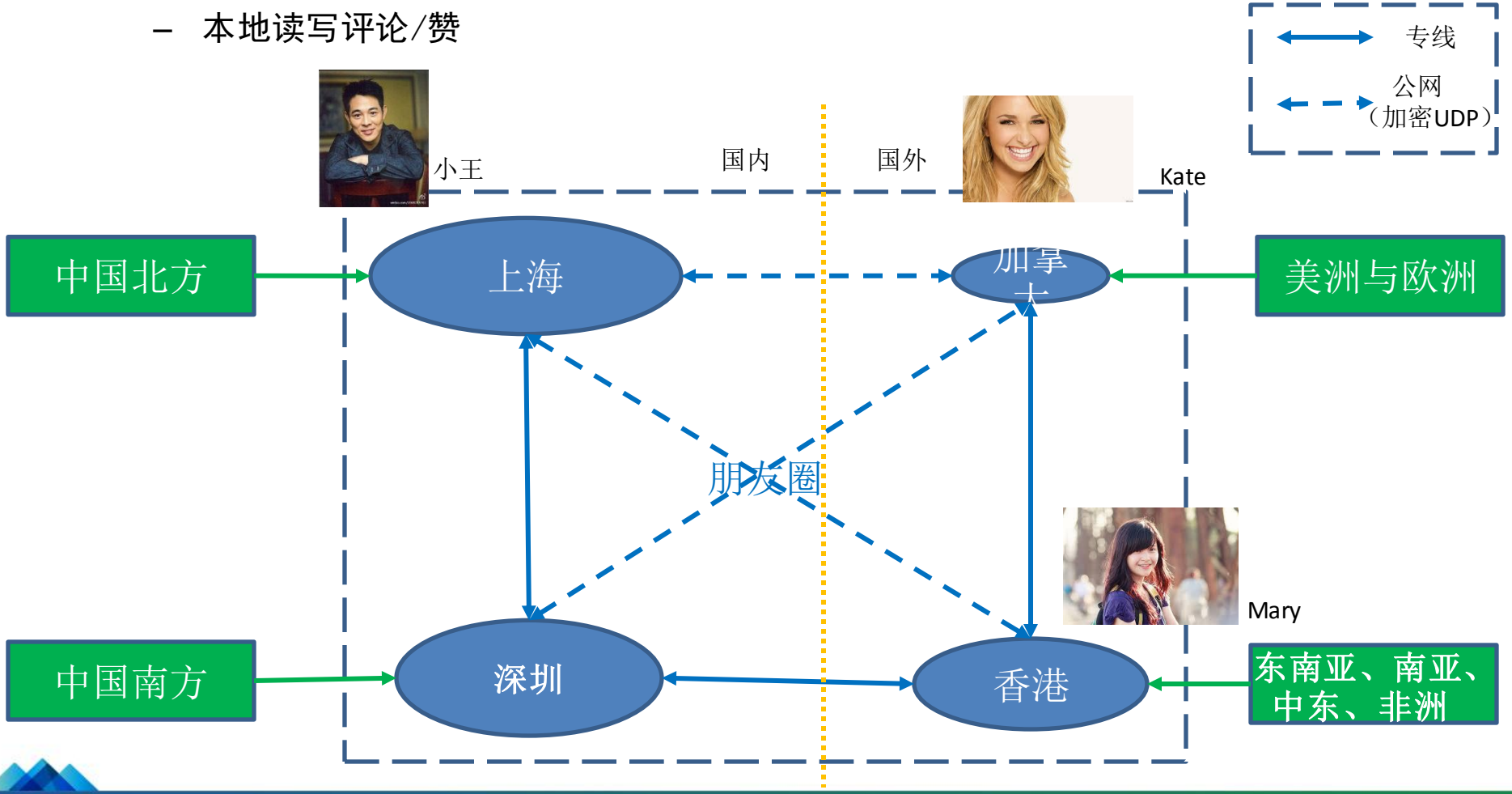


跨洋同步的挑战与应对

- 挑战
  - 大带宽高延迟
    - 国内 vs. 跨洋= 20-40 vs. 200-400ms RTT
  - 常见的丢包与乱序
  - 低可靠性：专线中断可预期
- 应对措施
  - 延迟、丢包、乱序达到一定程度时idc queue自动从TCP切换到UDP协议
  - 专线中断时自动切换到公网
    - AES加密

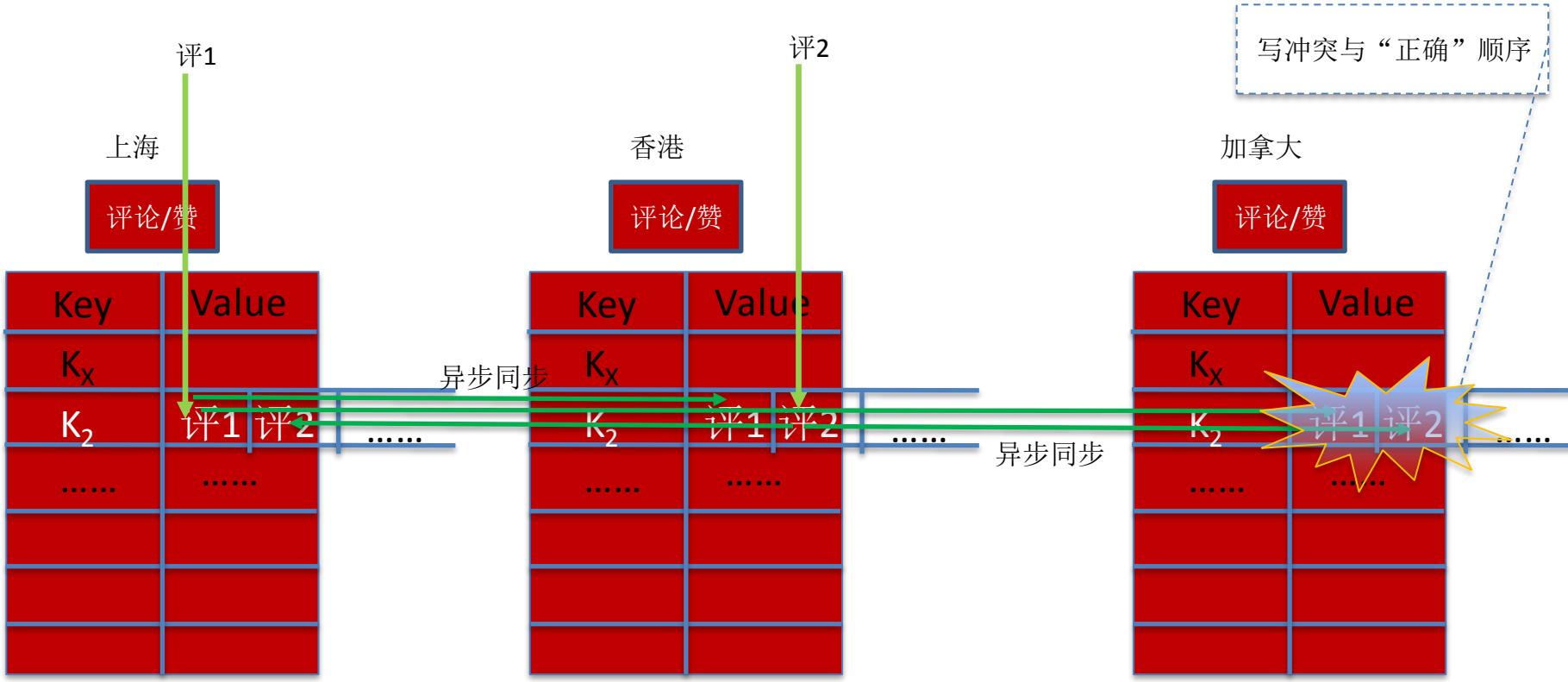
One More Thing

- 每个区域独立服务本区域朋友圈用户
  - 本地读写评论/赞

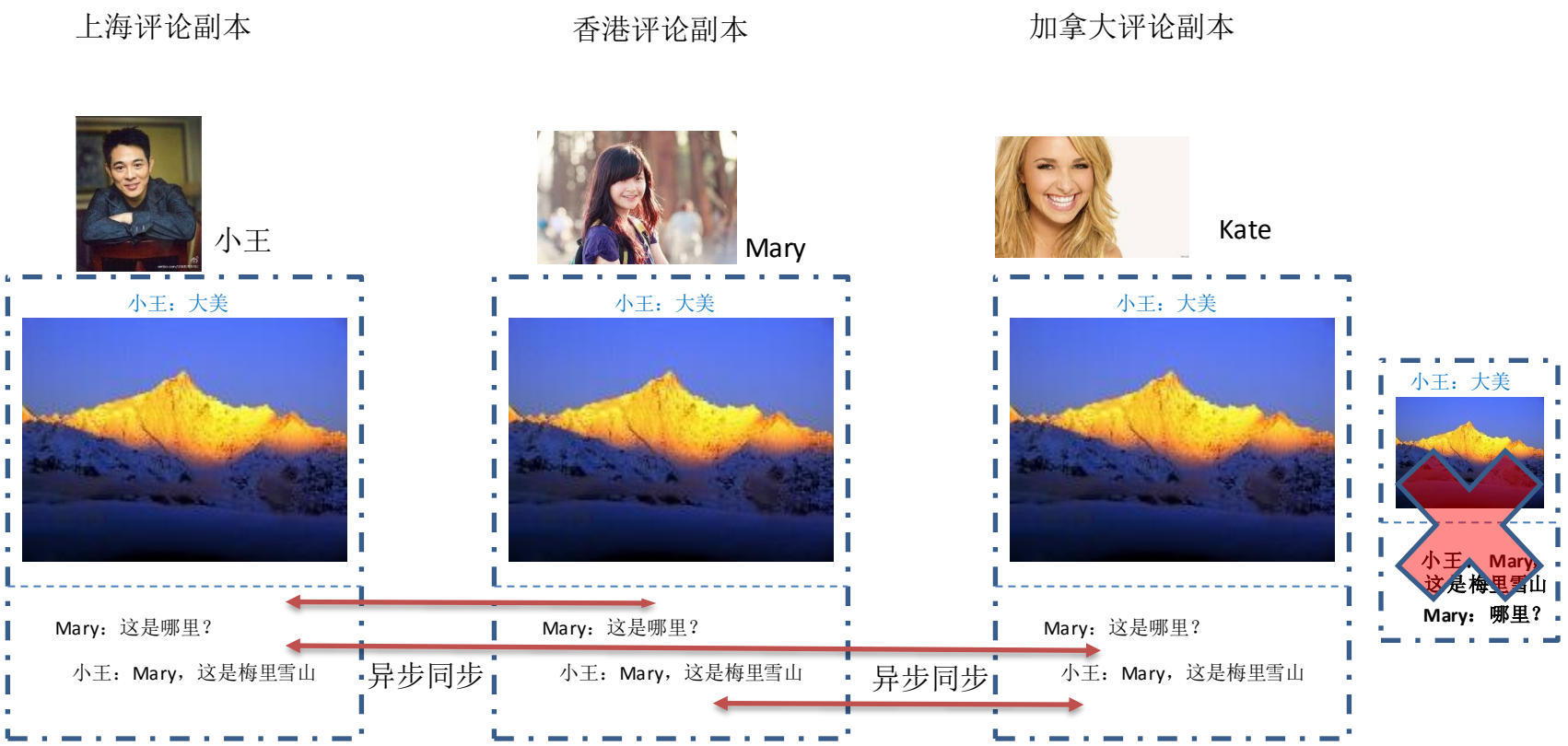


评论/赞的写冲突、“正确”顺序、与一致性

- 评论/赞
  - 本地写入
  - 异步双向同步到其它IDC



正确性与一致性

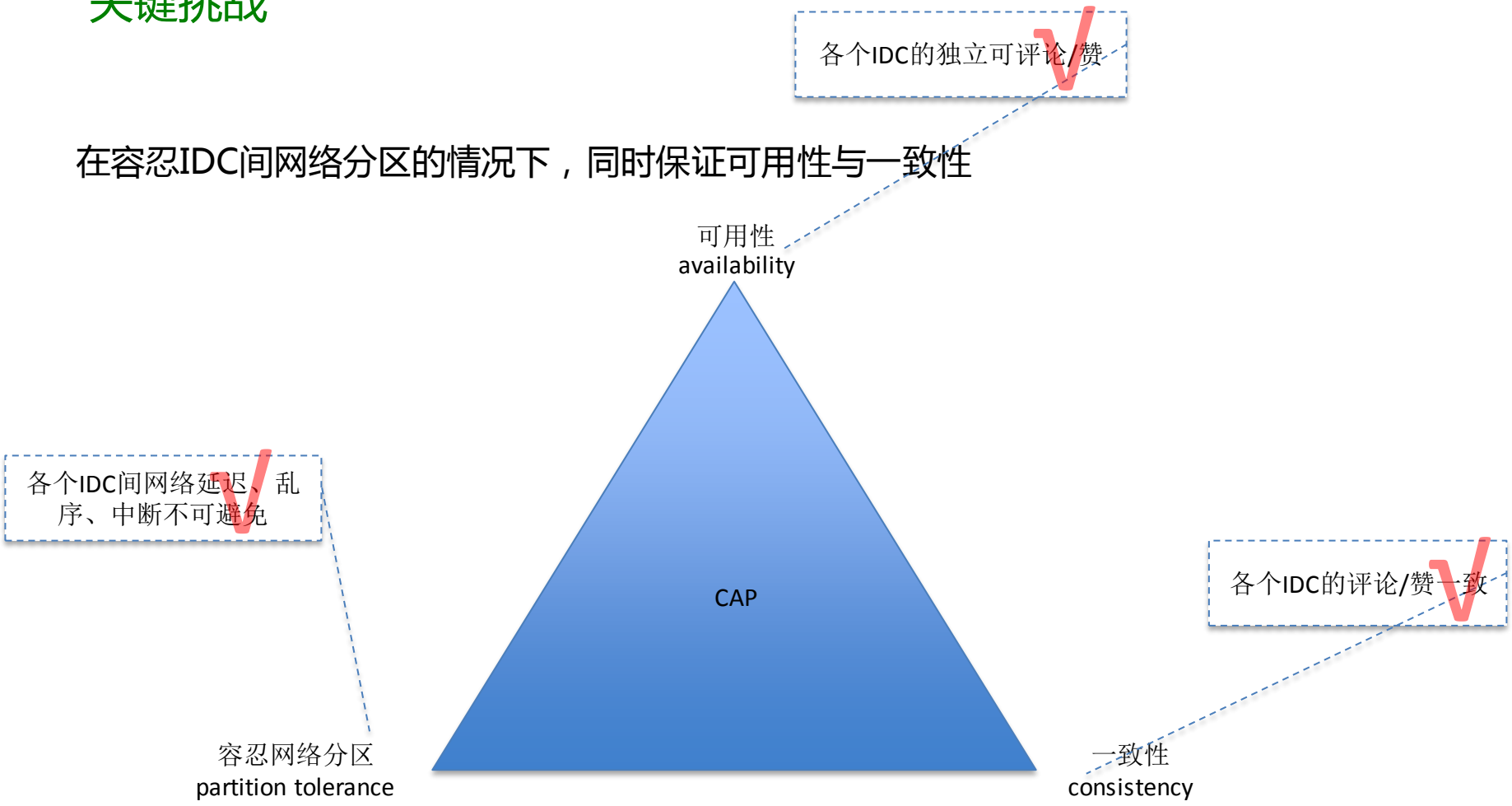

上海评论副本
香港评论副本
加拿大评论副本
小王
Mary
Kate
小王：大美
小王：大美
小王：大美
Mary：这是哪里？
小王：Mary，这是梅里雪山
异步同步
Mary：这是哪里？
小王：Mary，这是梅里雪山
异步同步
Mary：这是哪里？
小王：Mary，这是梅里雪山
小王：大美
小王：Mary
这是梅里雪山
Mary：哪里？

常见解决方案

1. 主-备同步
    1. 算法
        1. 每个key有一个主副本
        2. 评论/赞先写到主副本
        3. 主副本广播给次副本
    2. 优点：简单
    3. 缺点：存在单点失败的风险
2. 多点并行写入、按照评论/赞时间排序
    1. 算法
        1. 各个IDC本地写入、广播给其它IDC
        2. 客户端按照时间顺序展示评论/赞
    2. 优点：简单
    3. 缺点
        1. 分布式系统各个机器的时间不对齐、可能导致评论/赞乱序与相互覆盖
        2. 网络延迟可能导致前置评论暂时丢失
3. 中心锁
    1. 不予考虑

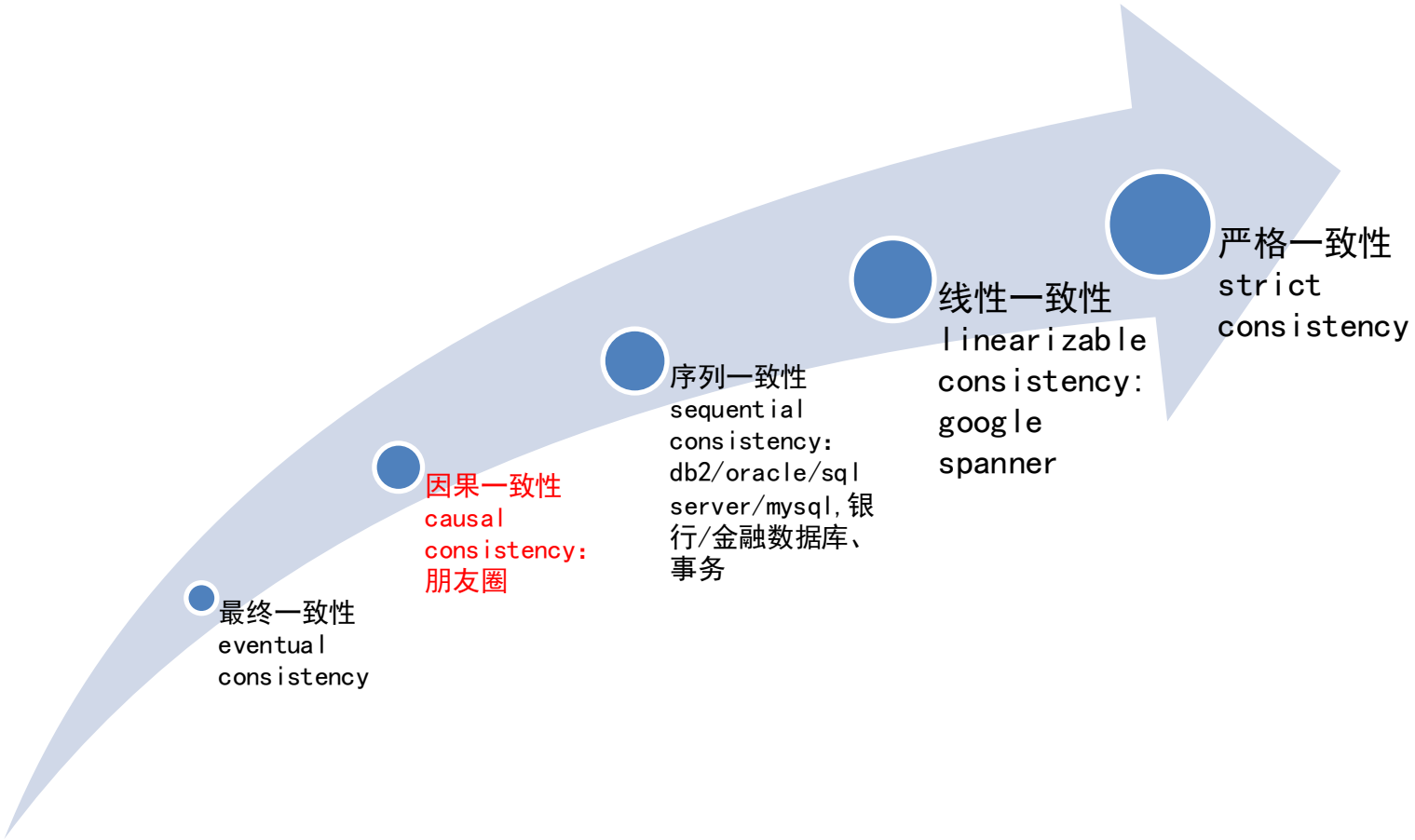
关键挑战

在容忍IDC间网络分区的情况下，同时保证可用性与一致性



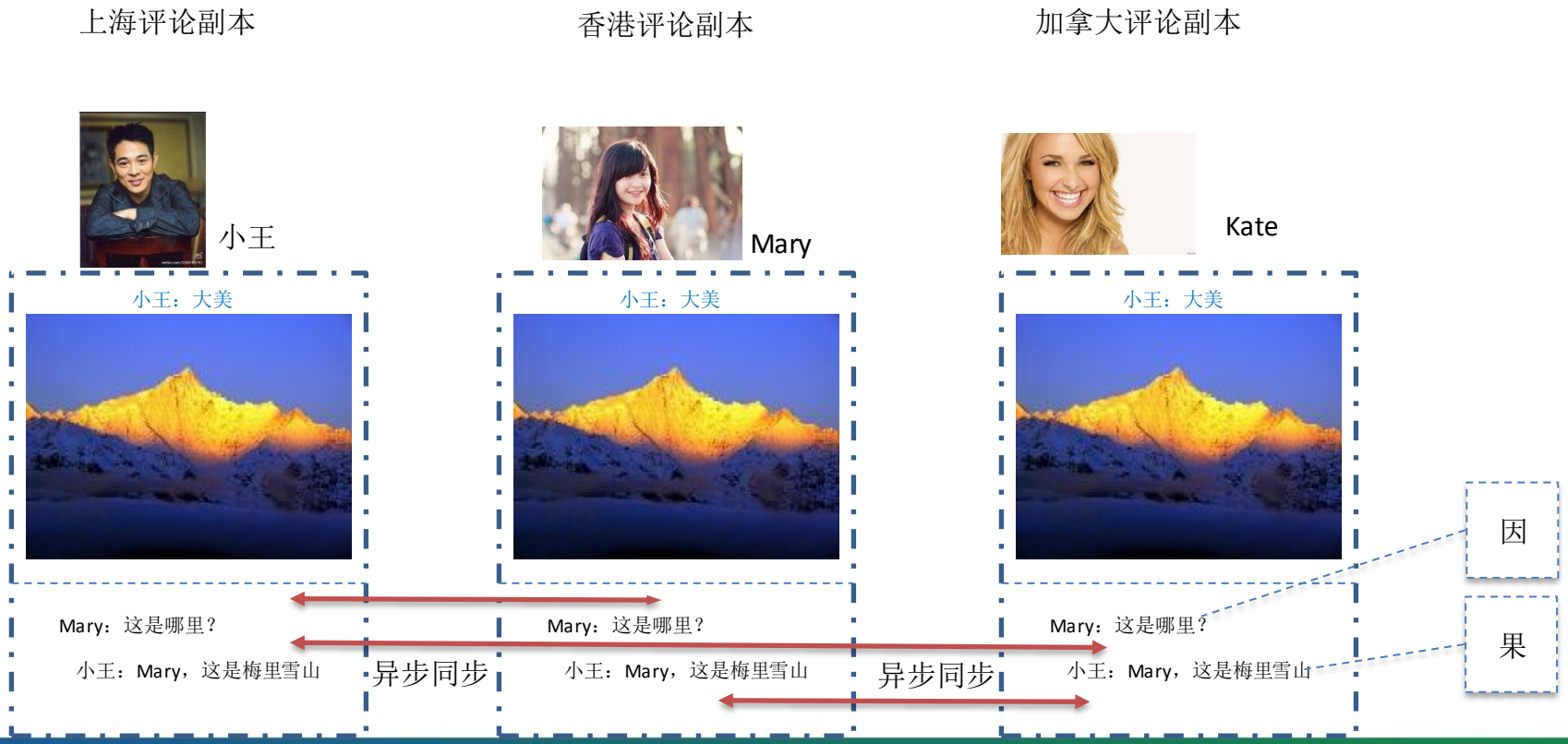
一致性

- 定义：一个对象的多个副本的数据一致

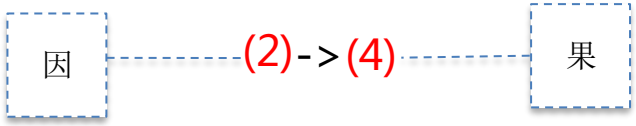


因果一致性

- 有因才有果
- 因发生在果之前
- 先看见因，才看见果
  - 先看见评论（因），才看见对此评论的回复（果）



定因果



上海评论副本　　香港评论副本　　加拿大评论副本

小王

Mary

Kate

小王：大美

(2) Mary：这是哪里？

(4) 小王：Mary，这是梅里雪山

异步同步

小王：大美

(2) Mary：这是哪里？

(4) 小王：Mary，这是梅里雪山

异步同步

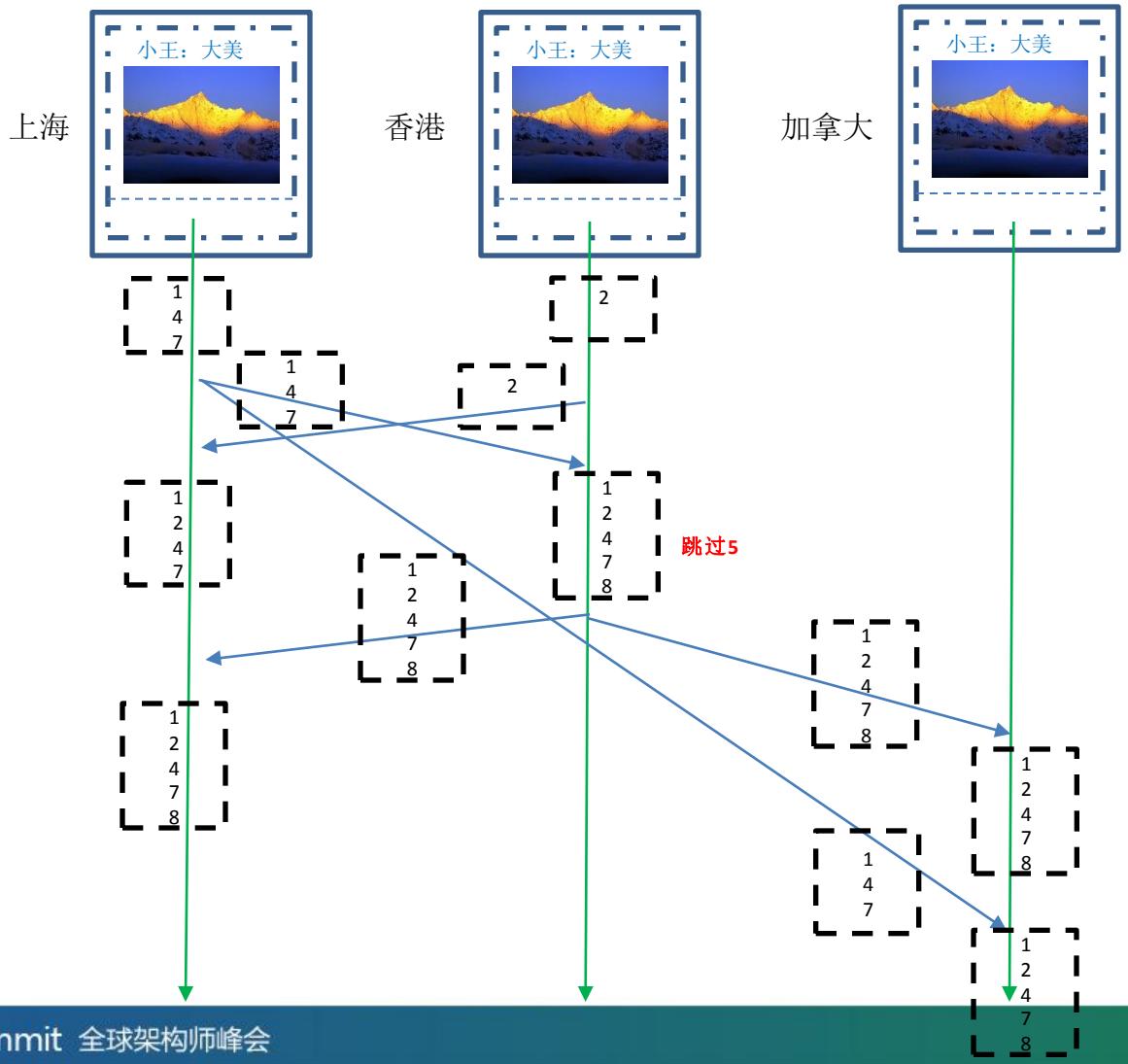
小王：大美

(2) Mary：这是哪里？

(4) 小王：Mary，这是梅里雪山

因果一致性算法

- 每条评论都有一个唯一的且递增的数字ID，确保排重
- 每条新评论的ID都必须比本地已经见过的全局最大的ID大，确保因果关系
- 广播本地看到的所有评论和新评论到其它IDC；相同ID的评论合并排重



1. Vector clock
2. Log-structure comments

Thanks!

朋友圈“微”团队诚邀第“4”人加入

- 如果你想踏上移动互联网的巅峰
- 如果你想服务全球亿万用户
- 如果你想挑战世界级的分布式系统技术难题
- 如果你想和一群志同道合的码农一起愉快玩耍

请加入我们！