十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店

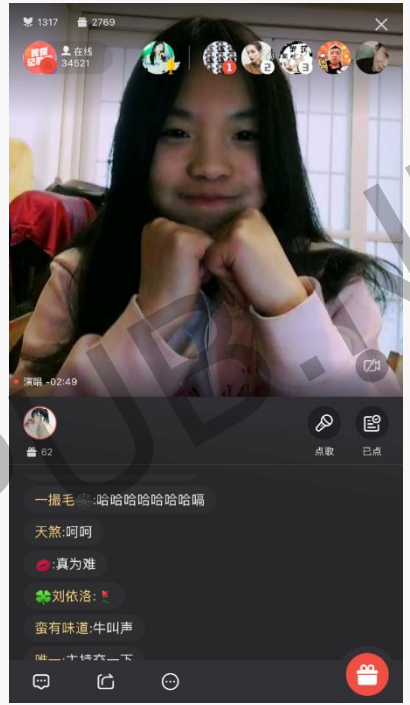
# 全民K歌直播消息服务

腾讯音乐 陈文武

# 自我介绍

- 2014年加入腾讯
- 核心基础服务负责人
  - 伴奏信息服务
  - 虚拟礼物体系
  - UGC动态服务
  - 直播消息服务

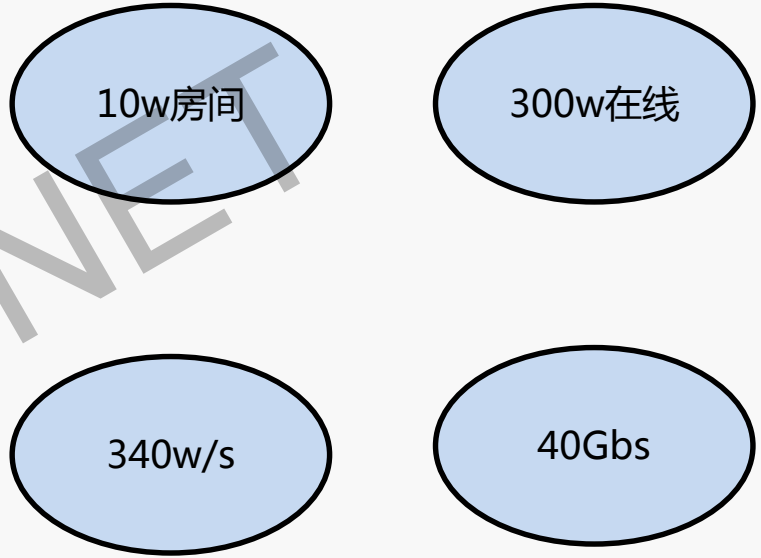
# 全民K歌直播与歌房


- DAU5000w+
- MAU 1.6亿+

# 海量消息挑战

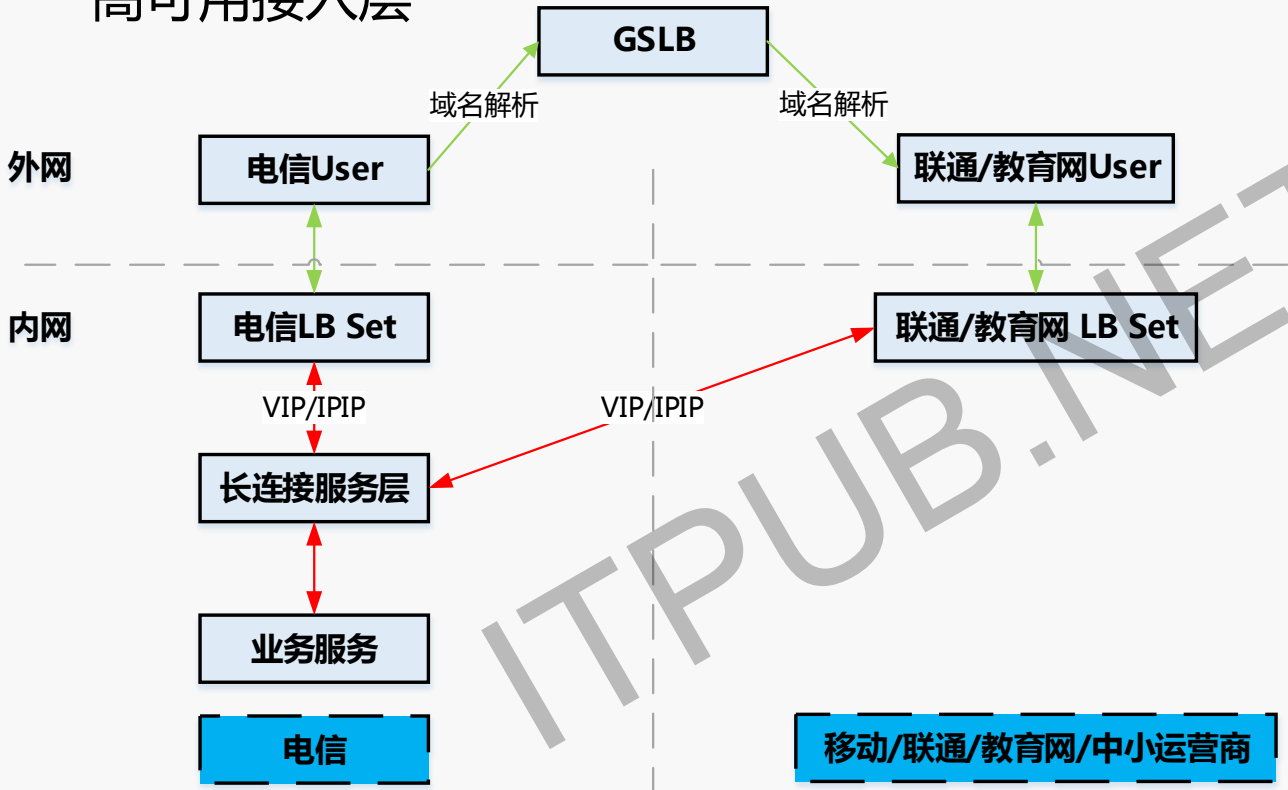
- 热点问题
  - 热门直播大量消息扩散带来巨大的带宽和负载压力
- 低时延
  - 互动消息延迟容忍性差
- 消息到达率
  - 关键消息要求可靠触达，很多关键操作依赖消息进行同步
- 不确定性
  - 互动引发的峰值，例如主播pk、礼包、运营活动等带来的在线人数的突发上涨
- 多地接入
  - 多地接入对流量和时延带来的挑战



# 巨人的肩膀

- 高可用网络接入层
- 成熟RPC框架
- 高可用KV存储系统

# 高可用接入层




**GSLB：域名解析优化**

- 省份城市，运营商最优解析
- 多地部署
- OSPF架构容灾，多台服务器一个VIP

**LB：内外网联通与负载均衡**

- 多运营商接入互联
- 防DOS
- IP收敛
- 4机一组，OSPF架构容灾，会话同步
- IPIP隧道模式屏蔽业务机器

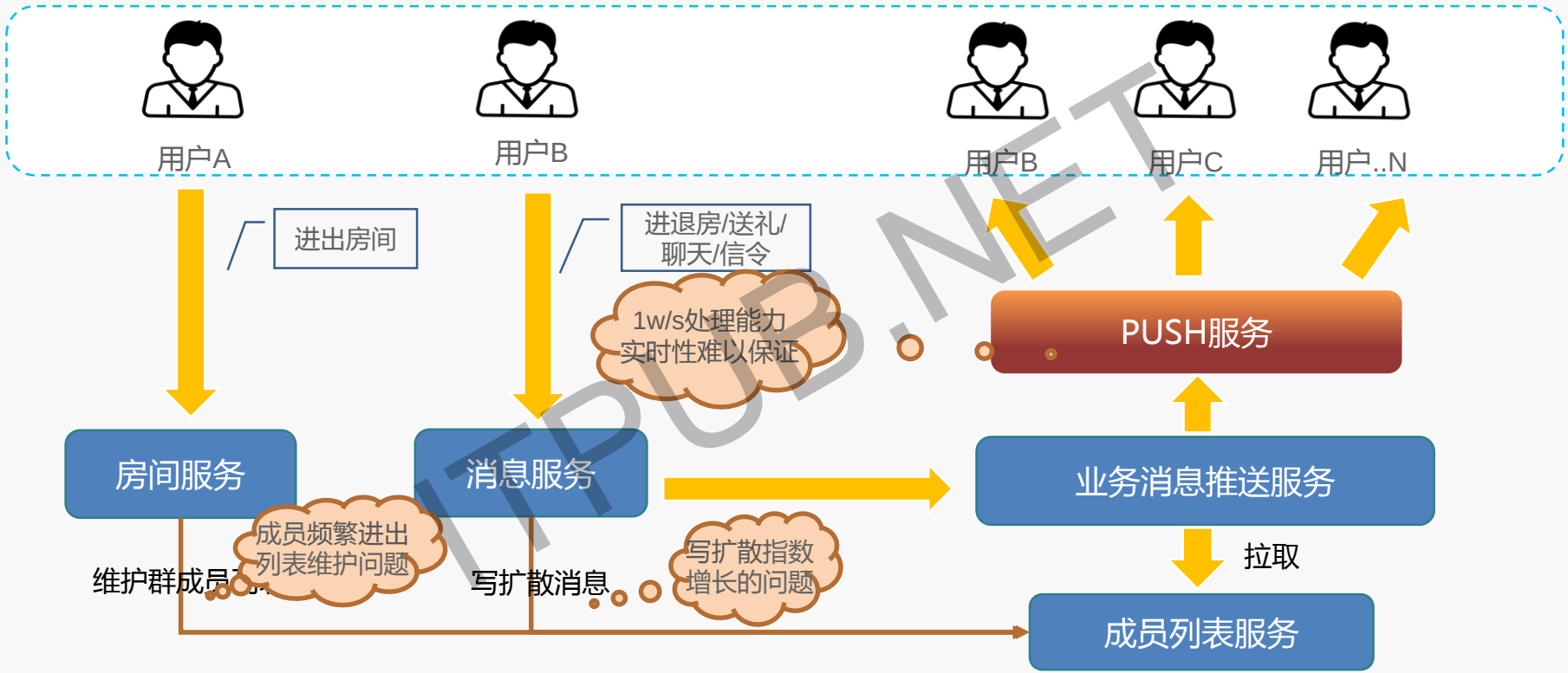
**长连接服务：连接通道优化**

- 加密通道
- 多地部署
- 接入调度与重定向
- 移动网络优化
- User端存IP列表，防止域名劫持，容灾

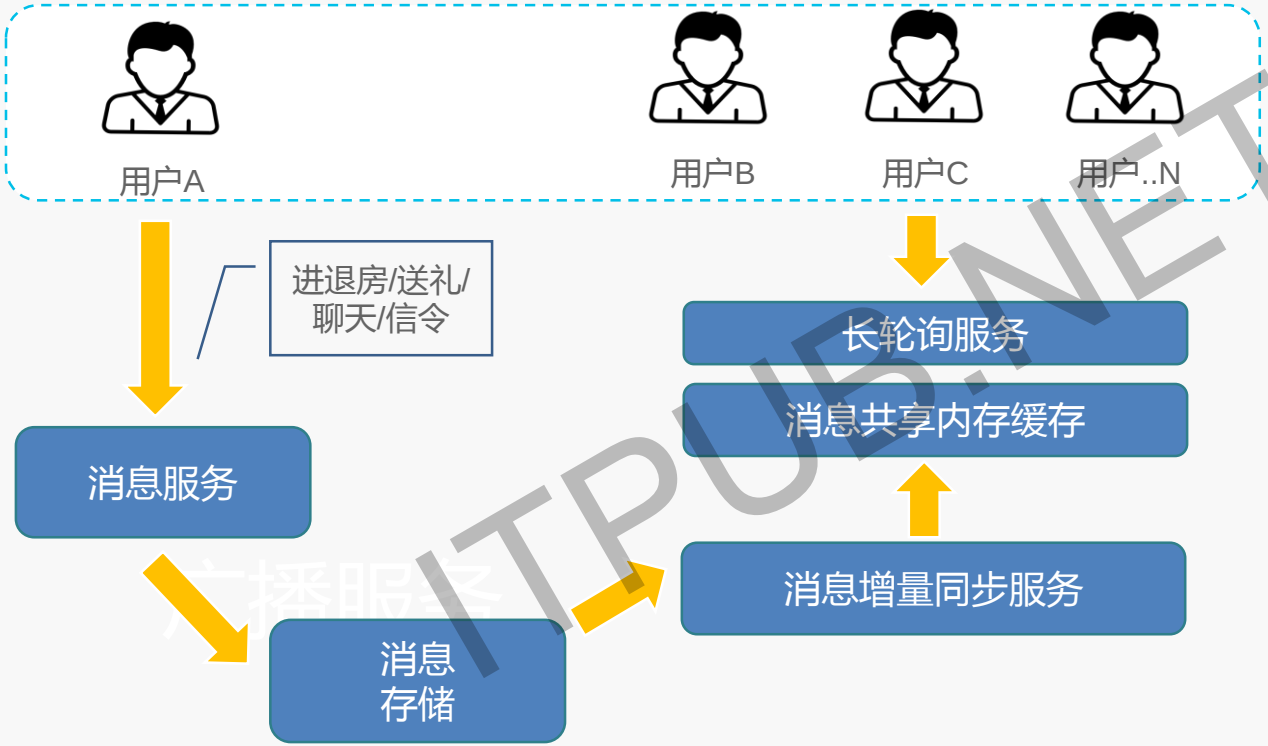
# 大纲

- 消息分发模型
- 关键设计
- 整体架构
- 读写容灾
- 过载保护
- 监控指标

# 消息分发模型—Push

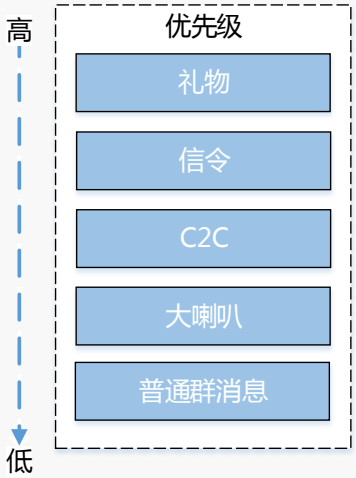


# 消息分发模型—Pull




- 实时性基本无区别
- 无须维护成员列表
- 没有写扩散
- 消息缓存减少流量穿越
- 实现简单可控性强

# 隔离设计




**消息存储隔离**

●消息按照不同优先级与类型，物理上隔离存储

●消息存储与其它业务物理隔离

**大小房间Set隔离**

●签约主播默认大房间

●非签约直播默认小房间

●满300人后迁移到大房间

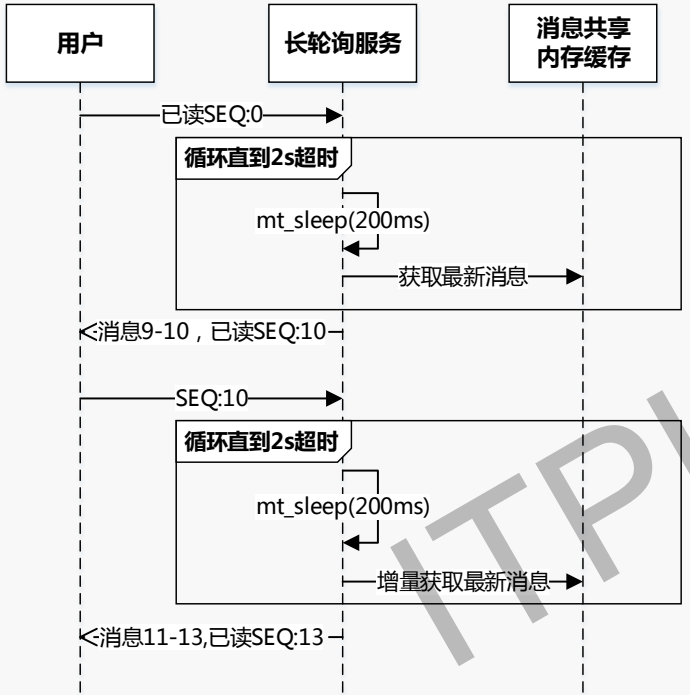
**K歌业务隔离**

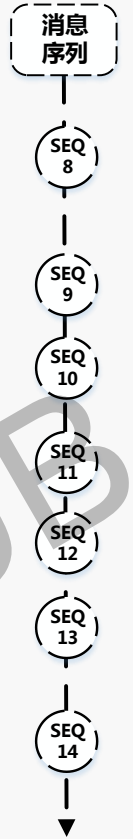
●独立分发层

●大房间无状态负载均衡路由

●小房间一致性Hash路由

# 消息SEQ机制



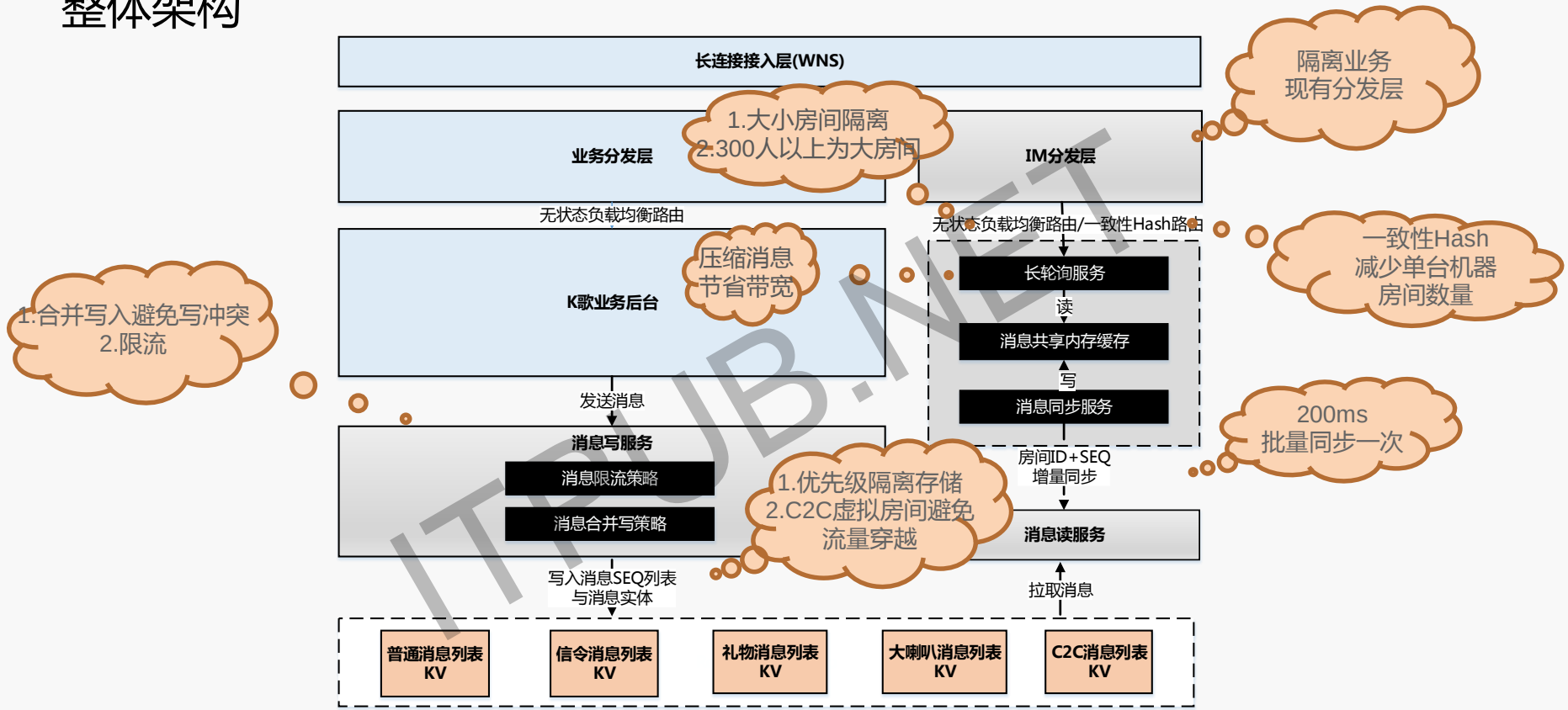

- 64位连续递增SEQ
- SEQ机制保证不重复读取
- 休眠200ms减少请求次数
- 内部轮询2s减少小房间的空查询(在调优调整)

大喇叭列表： seq=10
C2C消息列表：seq=2
群消息列表： seq=100
礼物消息列表：seq=20
信令消息列表：seq=50

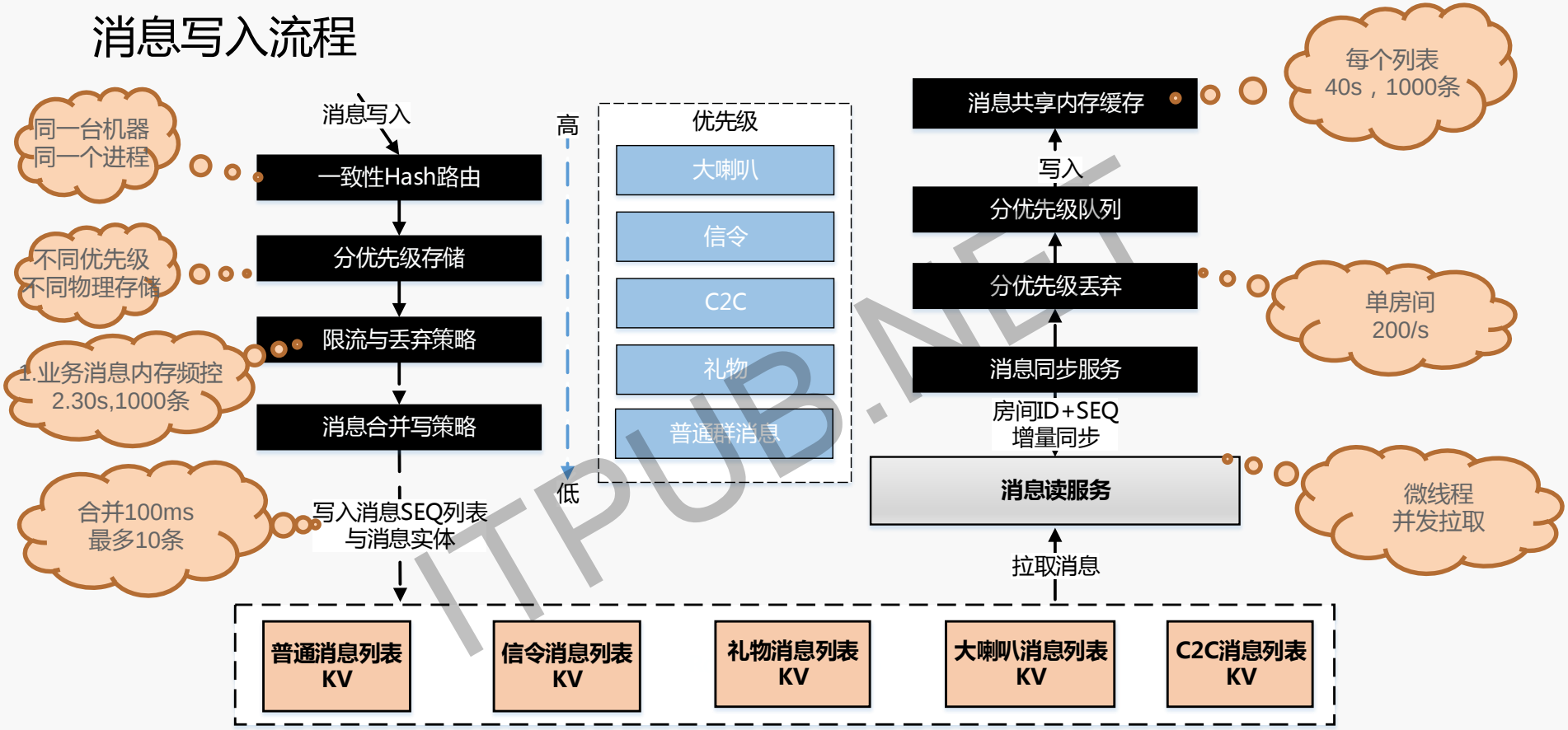
# 存储设计

- 存储结构
  - 存储消息SEQ索引列表，与消息实体数据
  - 只存最近30s，1000条消息
  - 30s？
  - 1000条？

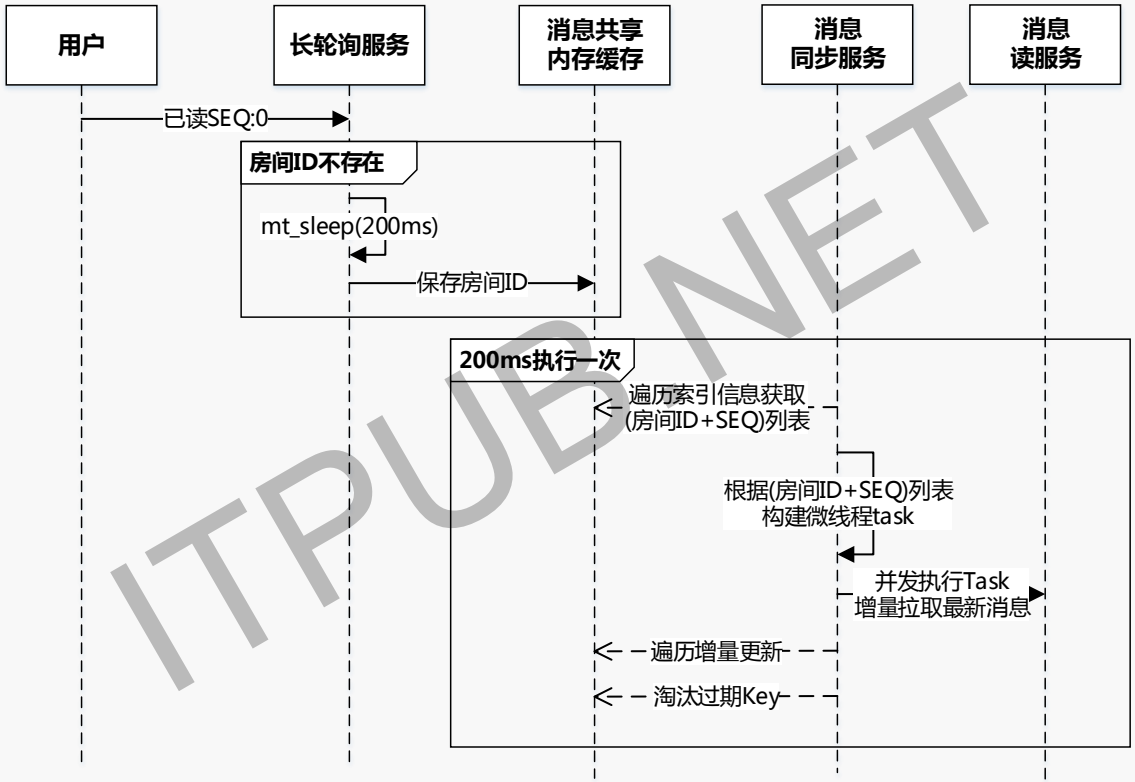
# 整体架构



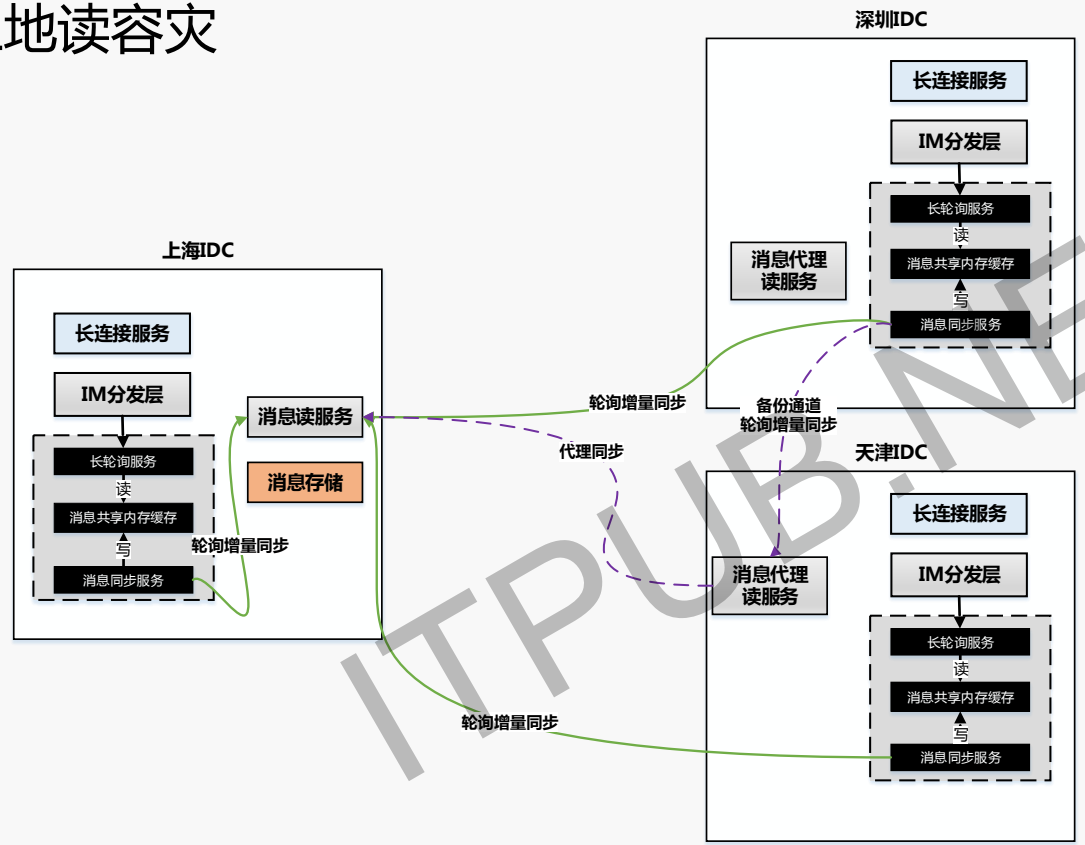
# 消息写入流程



# 增量更新流程

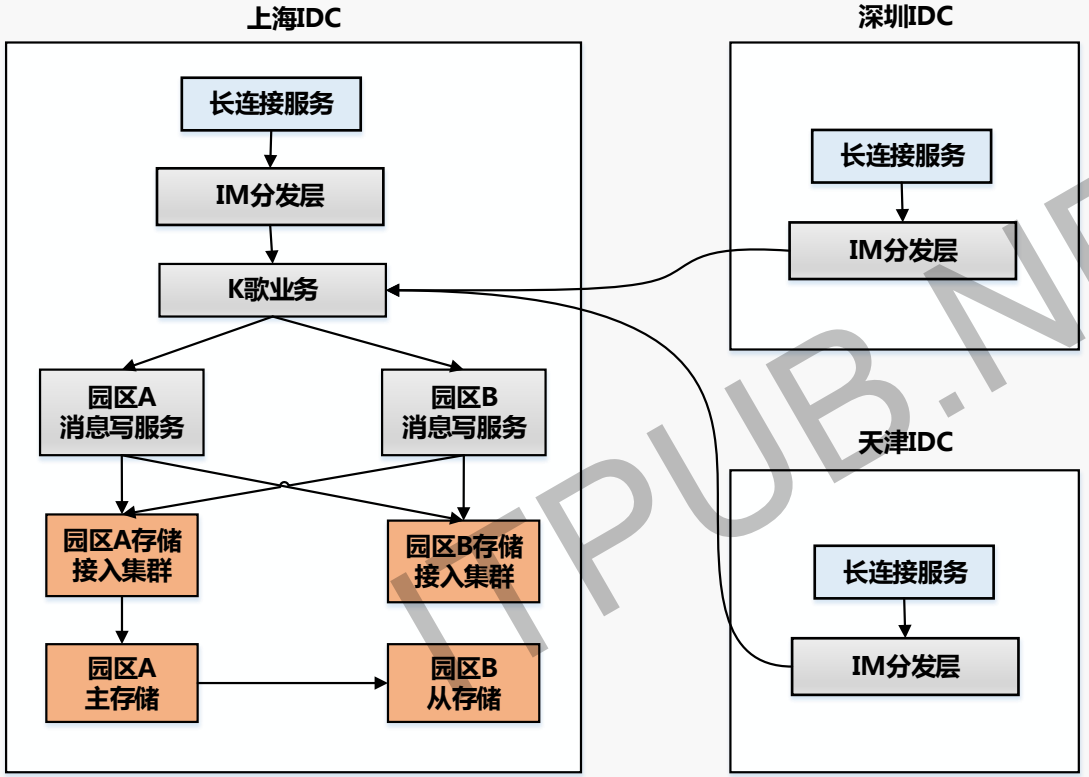


# 三地读容灾




1. 通过命令字隔离直播间
2. 三地部署，就近读取，避免跨城流量穿透（跨城IDC带宽成本高，时延高30ms）
3. 每一地的服务都为双园区部署
4. 天津，深圳专线互为备份(上半年3次专线故障)
5. 优化TCP慢启动过程，调优cwnd窗口(10MSS)
6. 减少单次同步消息量，优化跨城传输时延(同城55ms，跨城85ms)

# 上海双园区写容灾



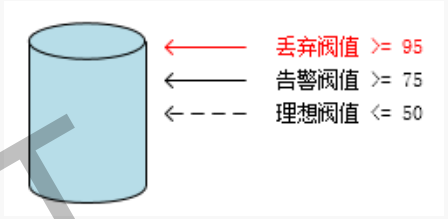

1. 每一地的服务都为双园区部署
2. 存储物理机主从分布在不同园区，故障秒级切换
3. 不同优先级消息列表，存储物理独立，并且互为备份

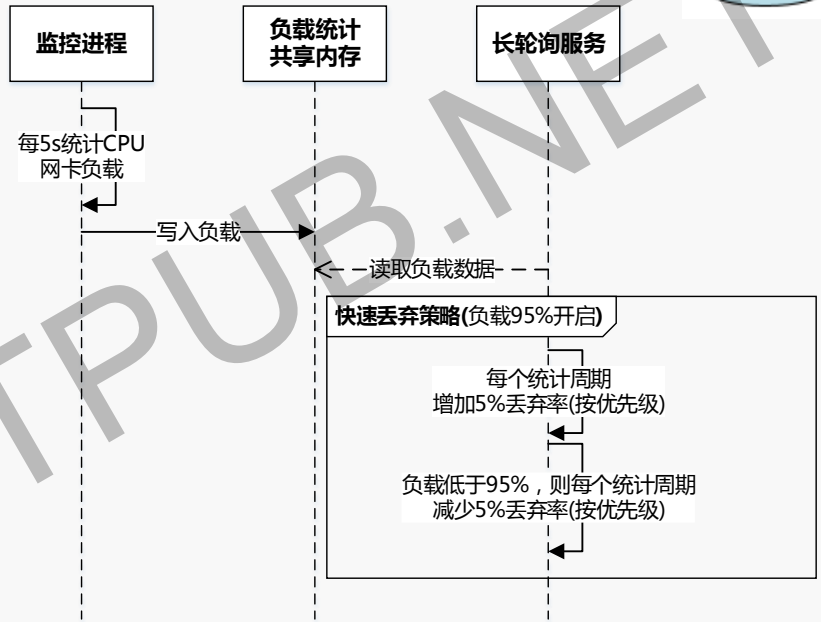
礼物消息列表：seq=20
信令消息列表：seq=50

礼物消息列表：seq=20
信令消息列表：seq=51

# 过载保护

- 负载均衡组件过载保护
- 服务端控制长轮询请求间隔
- 单次拉取消息上限保护—200条，优先读取高优先级消息
- 快速丢弃
  - 消息读快速丢弃策略



# 过载保护

- 最大处理能力：24000/s，400ms
- 请求超时时间2s
- 请求增加30%进入过载状态

**未开启快速丢弃**

处理能力：16000/s

处理时延：1.4s

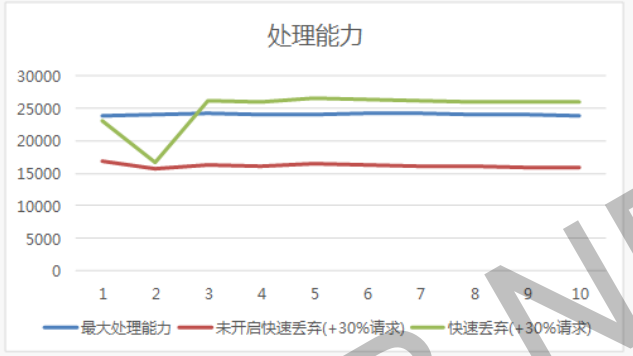
成功率：70%

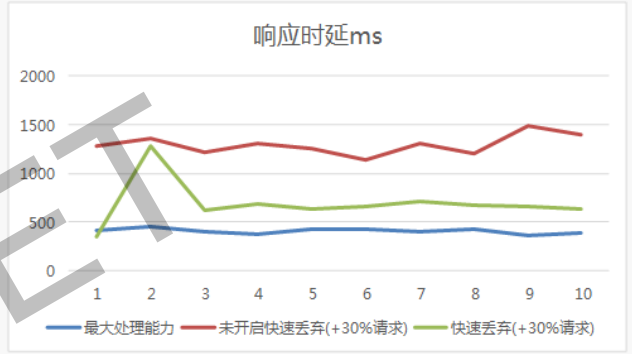
**快速丢弃**

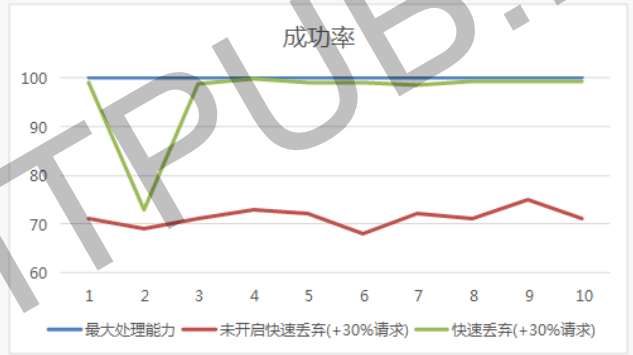
处理能力：26000/s

处理时延：650ms

成功率：99%



# 监控指标

- 长轮询SEQ连续性监控消息丢失率（99.99）
- 消息轮询读取延迟监控（超过5s上报失败）（99.98）
- 空查询率监控(大房间空查询率5%，小房间高峰期空查询率30%)
- 消息写入，消息丢弃，消息同步与重要角色读操作日志上报
- 消息产生速度超过50/s的预警
- 共享内存使用率监控、脏数据、错误数据监控和告警
- 机器负载，流量告警
- 数据统计和报表每天邮件输出

# 总结

- 热点问题
  - 业务隔离，大小房间隔离，存储分优先级物理隔离
- 低时延
  - 三地部署，就近接入，长轮询，压缩消息，消息缓存，虚拟C2C房间，跨城同步优化
- 消息到达率
  - SEQ机制，三地读容灾，双园区写容灾，消息列表存储互为备份，跨城IDC专线备份
- 不确定性
  - 合并写，限流，频控，柔性策略，快速丢弃过载保护
- 多地接入
  - GSLB就近接入，消息同步缓存减少跨IDC流量穿透，优化跨城请求大小与TCP慢启动

# 求贤若渴

- 全民K歌后台团队
- 全民K歌客户端团队
- 全民K歌国际化团队

# THANKS