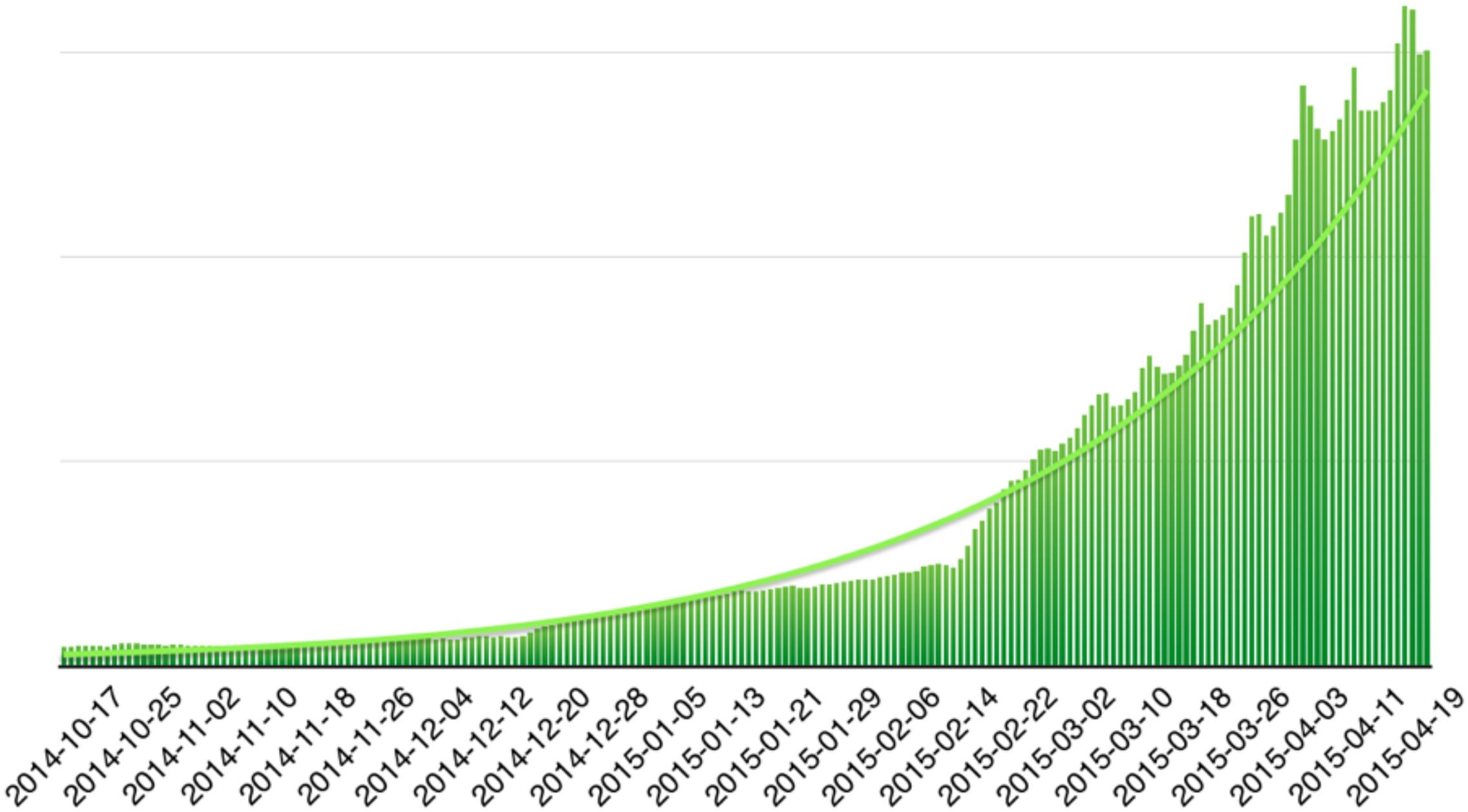
指数级增长业务下的服务架构改造

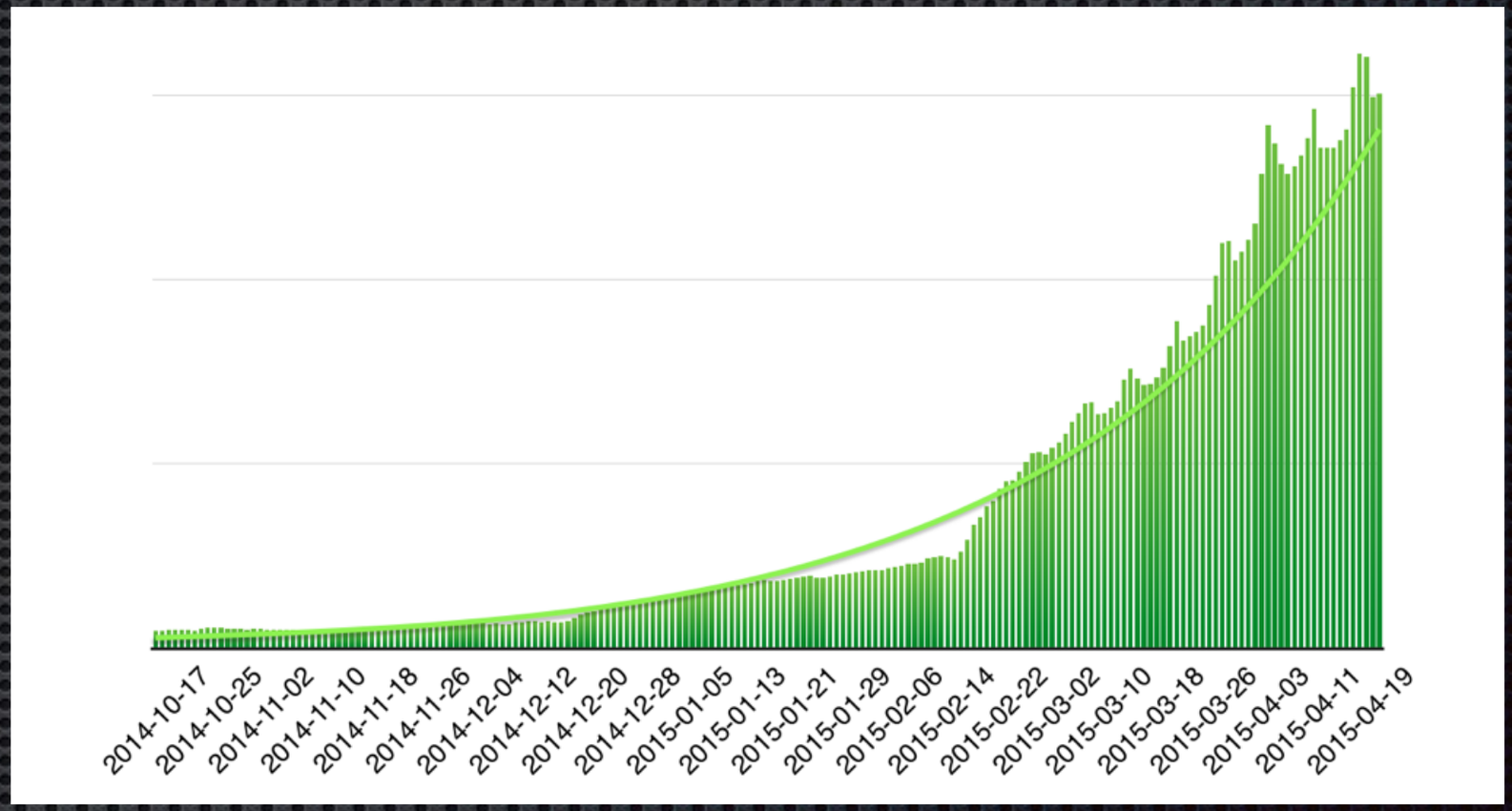
环信首席架构师
梁宇鹏 @一乐

增长速度


2014-10-17
2014-10-25
2014-11-02
2014-11-10
2014-11-18
2014-11-26
2014-12-04
2014-12-12
2014-12-20
2014-12-28
2015-01-05
2015-01-13
2015-01-21
2015-01-29
2015-02-06
2015-02-14
2015-02-22
2015-03-02
2015-03-10
2015-03-18
2015-03-26
2015-04-03
2015-04-11
2015-04-19

增长速度

- 每月一翻
- 春节假期 2x
- 同时在线近千万



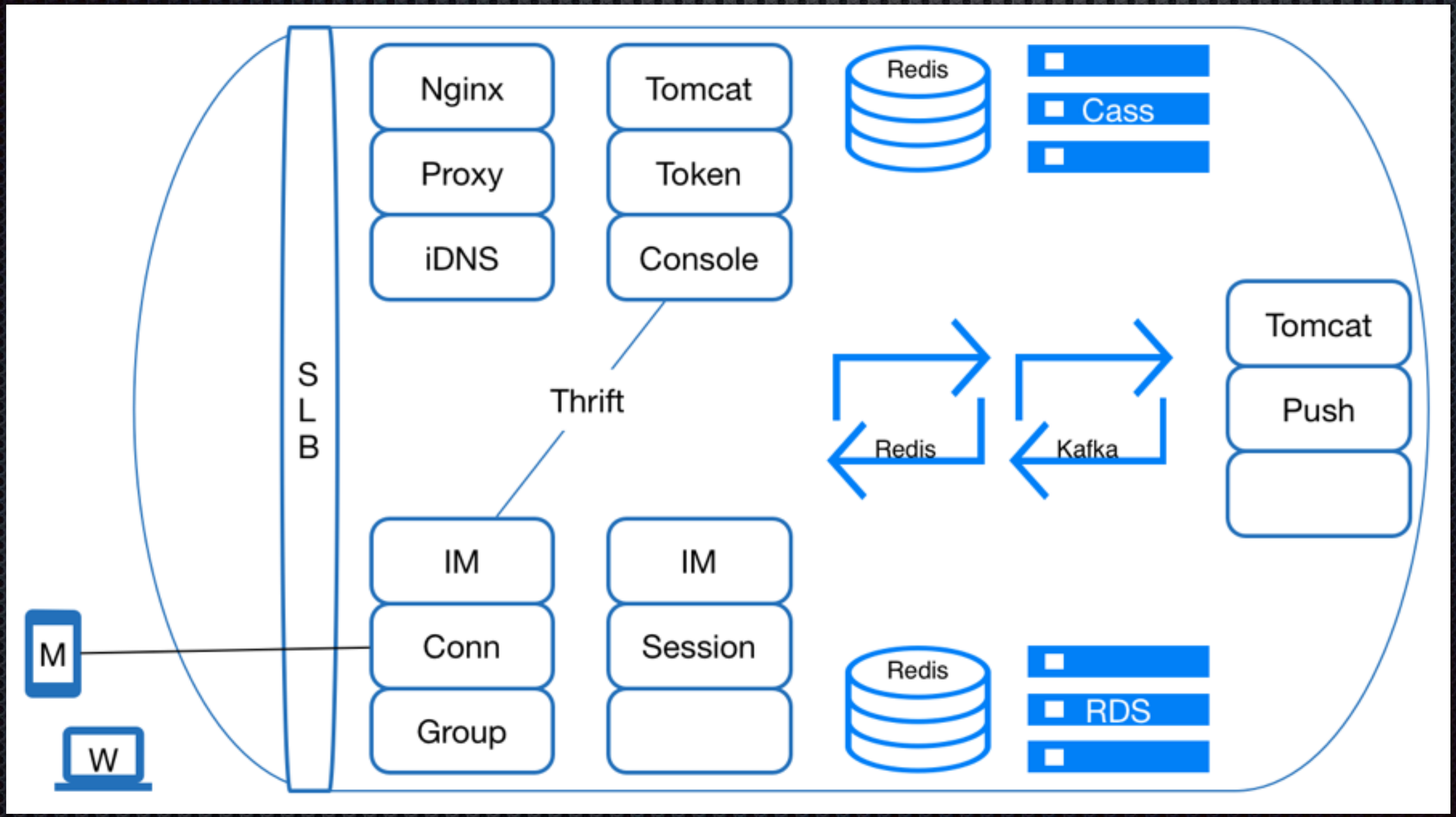
公司及个人

- 环信，即时通讯云服务
  - 帮助移动APP添加IM功能
- 专注IM领域
  - 环信首席架构师，原新浪微博通讯技术专家
  - XMPP开源项目Jabberd2、Ejabberd、Openfire

大纲

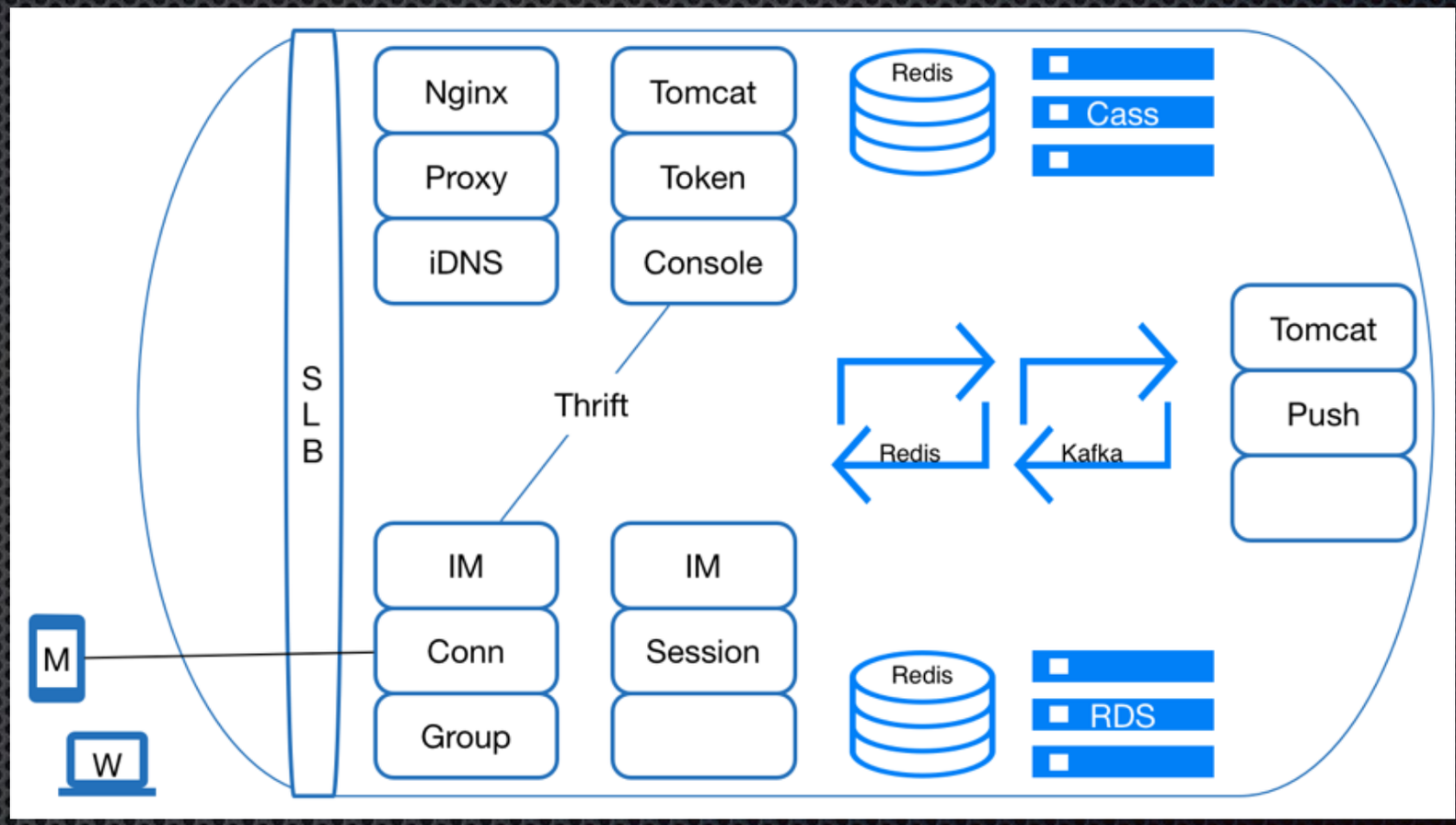
- 架构演化
- 经验教训
- 工具实践

服务架构


Nginx
Proxy
iDNS
Tomcat
Token
Console
Redis
Cass
S
L
B
Thrift
Redis
Kafka
Tomcat
Push
IM
Conn
Group
IM
Session
M
W
Redis
RDS

服务架构

- IDC 3
- Cluster Cell 4
- Machine 500+
- Erlang 120+



架构演化

- 100K -> 1M -> 10M -> more
- 伸缩性
- 可用性

伸缩性，通用

- 水平 -> 容量 -> 分区
  - MySQL，千库万表
  - Cassandra，动态扩容?
  - Mnesia，fragment
- 垂直 -> 性能
  - Redis缓存，业务独立

伸缩性，通讯

- 分层设计
  - 连接层、会话层、推送层
- 读写分离
  - 写库读缓 -> 文件服务，Token存储
  - 写处理迁移 -> 关系存储从IM到REST

可用性，通用

- 需求增加
  - 解耦，Kafka/Redis
- 峰值应对
  - 队列，群发消息流控
  - 降级，消息第一，登录次之

可用性，通讯

- 软实时
  - HOL blocking -> 队列迁移 or 清除
- 降级
  - 开关 -> 关停异常接口，留缓存舍DB
  - 数据恢复 -> 用户注册日志

大纲

- 架构演化
- 经验教训
- 工具实践

经验教训

- 不完美主义
  - 不多写代码 e.g. 会话存储拆分
- 头疼医头也医脚
  - 先容忍失败，再解决问题 e.g. 节点关闭逻辑
- 不头疼不医头
  - 量化分析 e.g. VM参数调整回滚

经验教训

- 未雨绸缪，超容量压测
- 峰值总比预期要早到来
- 数据仿真，从业务数据分析 e.g. 群用户分布
- 抓大放小 e.g. 登陆流程

经验教训

- 多租户多业务
  - 通用性外有易变性
    - 随新用户进来而改变
  - 意外热点
    - IN、TFboys

云上新挑战

- 性能波动导致承载容量下降
  - Noisy Neighbors
  - 取消数据节点磁盘快照
- 问题排查透明性?
  - 资源到服务，rds也要监控

云上新挑战

- 特定云问题
  - SLB心跳检查
  - SLB性能瓶颈
- 容量以及服务限制
  - 跨云设计
  - 只用公共特性

Dev & Ops

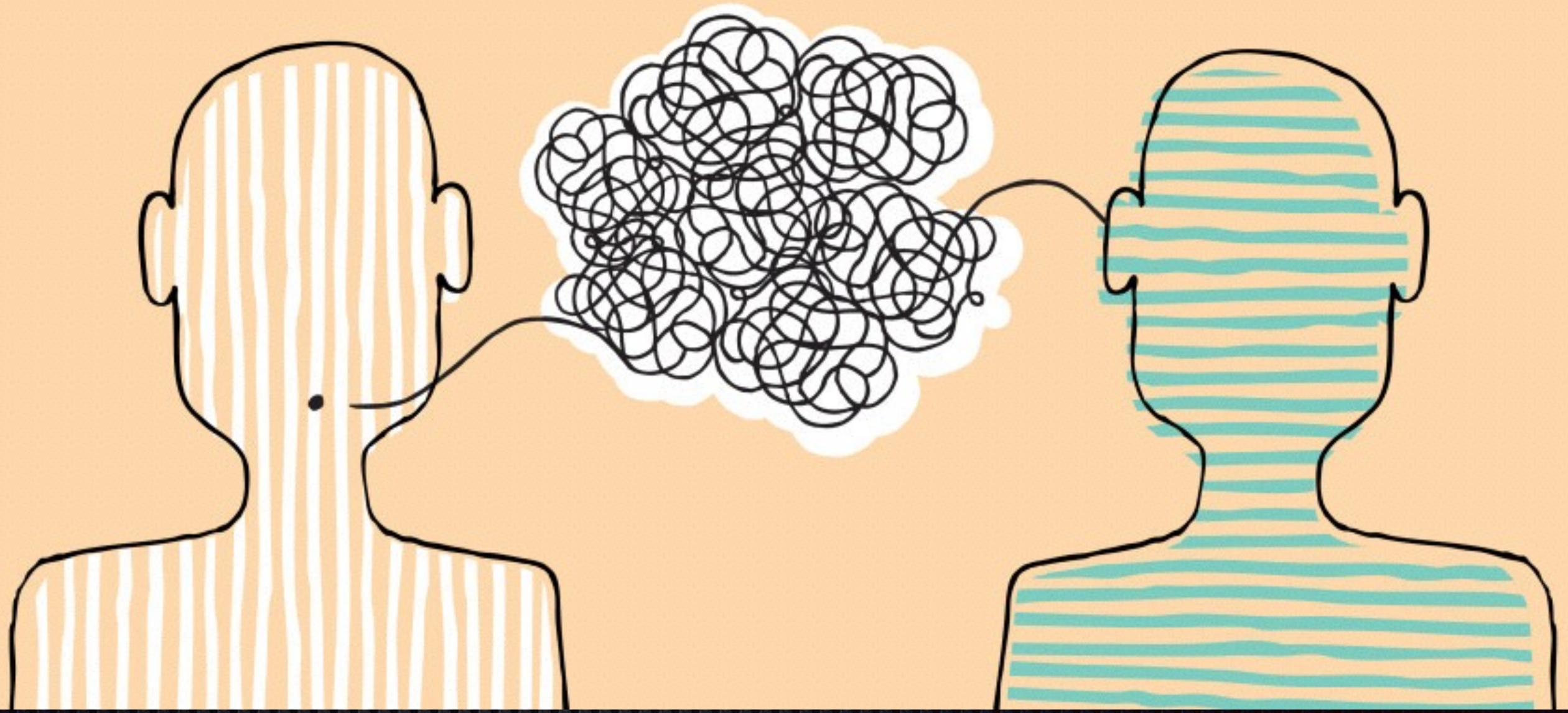
- Ops can dev 会武术有文化
  - Nginx 降级限流
- Dev for ops 好基友一辈子
  - 优化减少运维负担

演练

大纲

- 架构演化
- 经验教训
- 工具实践

语言问题



我们用Erlang

- 轻量级线程，消息传递
  - 并发友好，消息的软实时投递
- 速错 Fail Fast
  - 有错误不影响整体，健壮
- 不可变变量
  - 无副作用，不容易出错

一切皆缘

我们用Erlang

- 日志组件Lager
  - 同步异步切换问题，增加限流high_water_mark
  - flume backend + no thrift compact protocol support
- VM崩溃和调教
  - port_get_data in ERTS6.3 http://t.cn/RAOhWji
  - dist_buf_busy_limit: +zdbbl 2048000

我们用Erlang

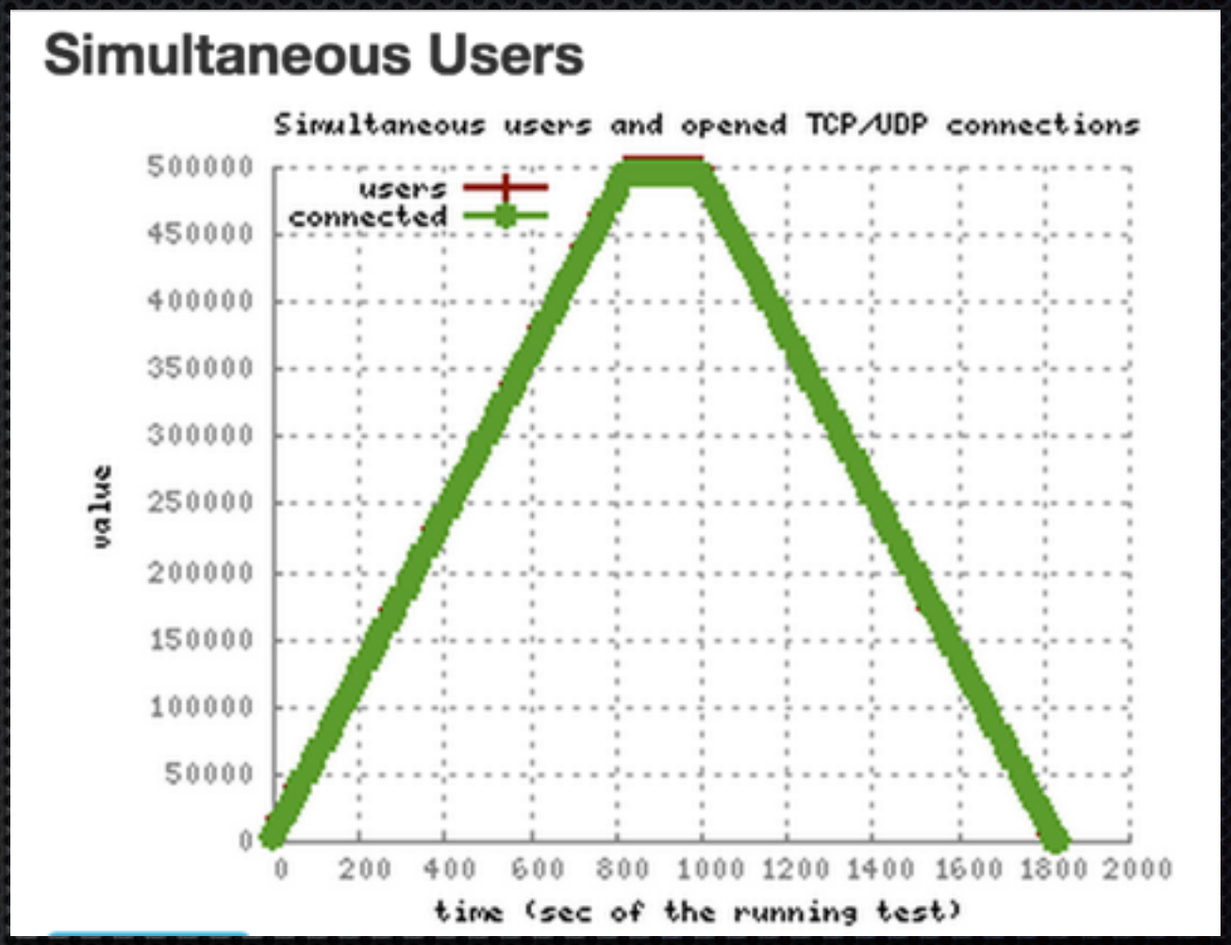
- 公平的调度器，并不公平的世界
  - 越来越多的线程
- Mnesia死等和数据改造
  - 多个节点同时重启更容易出现
  - 开箱即用 /= 开箱够用
  - new hash module for fragment

我们也用Java

- 语言像食物
  - 口味之争
  - 要开味蕾？要见世界！
- 语言只是工具
  - 一门语言就可以活
  - 不要为语言而活

压测工具

- TSung
  - TCPCopy
- 65535 problem?
  - ReuseAddr and Port
  - Client 200+



Transactions Statistics

Name	highest 10sec mean	lowest 10sec mean
tr_close1	0.675 msec	0.374 msec
tr_login1	18.16 msec	7.20 msec
tr_mucrooms	22.55 msec	4.95 msec
tr_rosters	15.56 msec	4.23 msec

团队协作

- 及时决策
  - 讨论要有定论
- 要自组织
  - 自我驱动，要有方向
- 远程办公
  - 注意沟通效率 + Slack Skype

未来展望

- 同时在线几千万、上亿
- 面向全球用户的即时通讯系统
- 微服务架构改造，Docker部署到开发
- 更多的优秀人才，不再2+3+4
- 多语言Erlang Java Go

谢谢

@环信即时通讯云

@一乐