IM架构分享

张彦

需求

- 基本的聊天功能
  - 点对点的聊天，支持收发文字、图片、语音、定位
  - 实时获得好友最新信息（例如：个性签名）
- 作为其他业务的基础设施
  - 触达：向特定的用户推送内容
  - 网络电话：信令交互

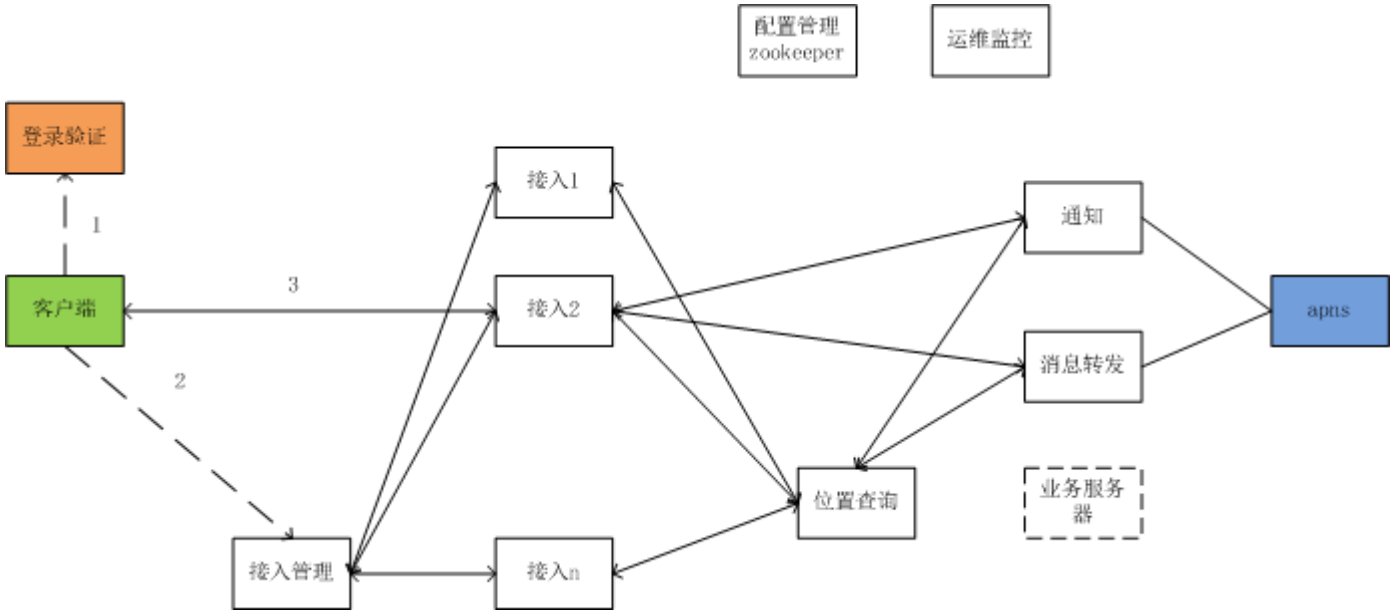
IM消息分类

- 即时消息
  - 若接收方不在线，可以丢弃
- 离线消息
  - 若接收方不在线，服务器要存储该消息，等到接收方上线后再发送
- 通知消息
  - 若用户在线，通知消息和即时消息一样
  - 若用户离线，同一类型的通知消息，服务器只需要存储最后一条

技术选型

- 自行开发
- 基于开源项目
  - openfire+spark搭建基于xmpp的im系统

逻辑结构图


配置管理
zookeeper
运维监控
登录验证
1
客户端
3
2
接入1
接入2
接入管理
接入n
通知
消息转发
位置查询
apns
业务服务
器

分层设计

- 接入层
  - 和客户端保持长连接
  - 根据业务码和功能码，转发客户端信息到后端业务服务器
  - 将服务器信息发送到客户端
- 业务层
  - 支持聊天业务
  - 支持其他业务

组件说明

- 登录验证服务器
  - 提供http服务接口
  - 密码验证，生产token
    - 最初的Token是和uid对应的一个随机字符串，登陆验证服务器还需要提供token验证的接口
    - 由于token验证的并发量巨大，后期token修改为一个加密串，接入服务器通过对token解密来验证，减轻了登陆验证服务器的压力

组件说明

- 接入管理服务器
  - 监控接入服务器状态，主要是负载情况
  - 根据用户所在区域及运营傻瓜，选择最优的接入点
  - 接入管理服务器通常会返回3个接入服务器地址，客户端选择最快的一个连接

组件说明

- 接入服务器
  - 维护与客户端之间的长连接，线路空闲超过5分钟会主动断开连接
  - 转发上下行的消息
  - 在消息头添加接入服务器编号
  - 通过消息队列
    - 向位置查询服务器发送上线/离线通知
    - 向接入管理服务器发送负载报告

组件说明

- 位置查询服务器
  - 向业务层提供用户位置和状态查询，包括用户是否在线，若在线是连接到哪一台接入服务器，用户的网络状态，手机平台信息，客户端版本信息，等
  - 向订阅了上线/离线通知的业务服务器发送相应的通知

组件说明

- 消息转发服务器
  - 即时消息转发服务器
  - 离线消息转发服务器
- 若消息头未记录服务器编号（或编号不符），会向发送方返回服务器编号
- 离线消息保存到mysql数据库，并通过redis缓存加速
- 订阅上线通知

组件说明

- 通知服务器
  - 存储/发送通知消息，和消息转发服务器类似，对离线用户的通知也分为可丢弃和服务器存储2类
  - 离线通知，支持同类型合并

组件说明

- 业务服务器
  - 基于im提供的能力完成其他的业务
  - 例如：网络电话、用户触达

组件说明

- 配置管理
  - 配置参数
  - 监控各服务器运行状态
  - 新服务器注册

组件说明

- 配置管理
  - 基于zookeeper
  - 存储配置参数
  - 新服务器注册
  - 监控各服务器运行状态，判断一个服务器是否不可用，还要结合心跳测试

组件说明

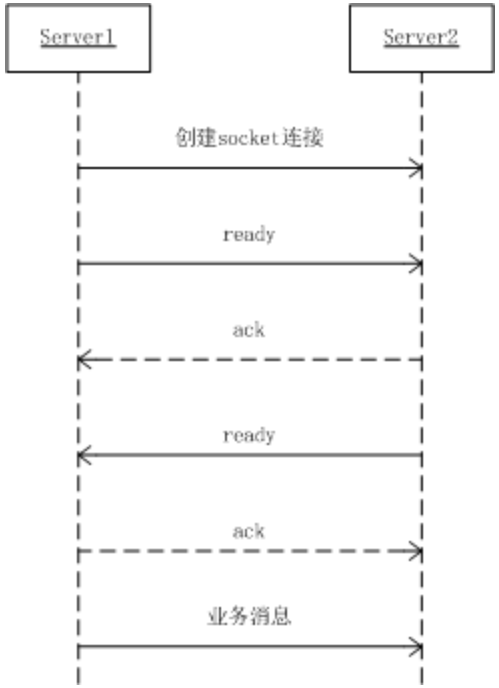
- 运维监控
  - 收集各组件的实时和统计信息，并通过图表的形式进行展示的后台系统

协议设计

- 报文头：bson格式
- 报文体：json格式
- 后续为了提升网络电话业务质量，引入了protobuf

服务器注册

- 首先向配置管理服务器注册，等待其他服务器来连接
- 连接后，要确认是否ready，双方都ready才能发送消息



服务器注销

- 注销服务时，需要先在配置服务器注销，然后向所有链接的服务器发送not ready信令
- 所有服务器都返回ack后，该服务器可重启

一些难点

- 客户端网络切换
- 消息排序
- 跨域部署
- 消息延时处理
- 网络抖动

客户端网络切换

- 关于上下线
  - 若用户断线后在极短时间内重连成功，且连接到同一个接入服务器，可不通知业务服务器上下线
  - 但是，若业务服务器对用户网络状态敏感，则需要特殊处理
    - 例如，网络电话业务并不把2G连接视同在线

客户端网络切换

- 客户端网络切换，根据重连的时间间隔，连接到的接入服务器和用户网络状态，位置服务器对业务层发送的通知不同
  - 离线通知延时发送，等待用户重连
  - 重连时间短，连入同一个服务器，网络状态相似（例如，多数时候3g和4g可认为相似），可以不发送通知

跨域部署

- 域是个逻辑概念，一个域是可以跨越多个机房的
- 在一个域内，多机房间采用专线互通
  - 位置查询服务器彼此是独立的，各自拥有自己的存储，其数据来源于消息队列
  - 消息转发服务器/通知服务器，存在一个集中的写入数据库，异地机房部署的服务器读写分离，读本地的从库，写入异地的主库
- 跨域通过跨域网关来实现互通，类似于电话网的分区，通过区号来区分用户所在的区

消息排序

- 客户端发送消息时要按顺序编号，编号的起点是连接到接入服务器时获得的
- 在一次会话中，客户端的消息，原则上由同一个消息服务器处理
- 接收方需要按顺序号调整显示效果

消息延时处理

- lm系统也许对延时不敏感，但电话系统对延时很敏感
- 为了减少延时，一个主要的技术手段就是减少消息包的尺寸，经过测试，单个包不超过500字节为好
- 用protobuf替换了bson/json格式

几种流行编码格式对比

对比测试结果如下

消息类型					压缩效率		
类型	子类型	名称	方向	请求/应答	目前长度	最大压缩长度	压缩比
Call	Call_req	呼叫请求	上行	请求	357	142	60.2%
Call	Call_req_ack	VPS 收到呼叫请求的 ACK	下行	请求	122	43	64.8%
Call	Ring_rsp	响铃响应	下行	请求	239	127	46.9%
Call	Call_rsp	呼叫应答	下行	请求	251	136	45.8%
Call	Call_ack	呼叫应答确认	上行	请求	126	76	39.7%
Call	Bye_req	结束呼叫	上行	请求	112	65	42.0%

消息类型		报文头编码长度					报文体编码长度			
类型	子类型	BSON	JSON	MsgPack	TLV	protobuf	JSON	MsgPack	bzip2	protobuf
Call	Call_req	54	58	30	21	13	299	206	210	125
Call	Call_req_ack	118	134	71	39		0	0	0	
Call	Ring_rsp	118	133	71	42		117	81	131	
Call	Call_rsp	118	133	71	42		129	90	136	
Call	Call_ack	64	71	37	24		58	48	90	
Call	Bye_req	64	71	37	24		44	37	82	

网络抖动

- 接入服务器崩溃，则所有用户需要重新连接，位置服务器要发送离线通知到所有订阅的业务服务器；若实际上接入服务器未崩溃，则又需要发送大量的上线通知
- 为了避免网络抖动造成的订阅消息爆炸，需要能识别抖动
- 当心跳测试超时或未收到心跳时，不会立即判断服务器不可用，会发送测试消息并重试3次，若测试消息全部超时，才会判定为服务器不可用

谢谢