OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识共享平台

2017源创会年终盛典

与电子标准院共建开源标准

12月23日　北京万豪酒店

主办方 开源中国 oschina.net

CESI

腾讯信鸽海量移动推送服务构建

数据平台部 甘恒通

1 推送系统建设

- 终端
- 基础设施
- 后台
- 云化治理

2 增值服务

- 精准推送流程
- 数据积累
- 支撑平台
- 可视化操作

移动推送服务概览

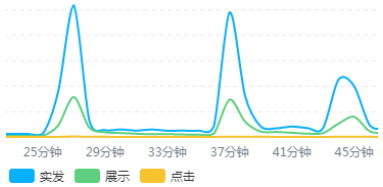
终端
- 共享 Service
- 灰度 可运营

基础设施
- 全球网络
- 自建 DNS
- 统一接入

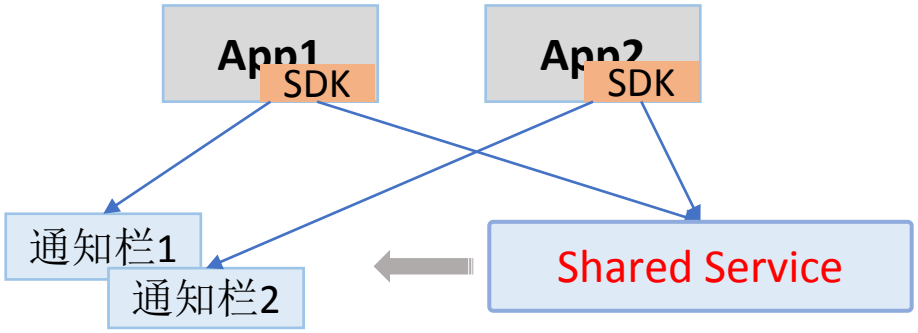
推送 SaaS
- 推送
- 路由
- 检索

实时效果跟踪、多维运营分析

增值服务
- 用户分群
- 精准推送
- 场景推送
- 泛推送
- 再营销



终端共享通道、服务可运营



功能模块

设备识别 | 通道集成 | 消息管理 | 效果统计

基础组件

安全审计 | 网络存储 | 质量监控 | 云控模块

正效果

展示点击

打开应用

负效果

关闭通知

应用卸载

App粒度云端调度

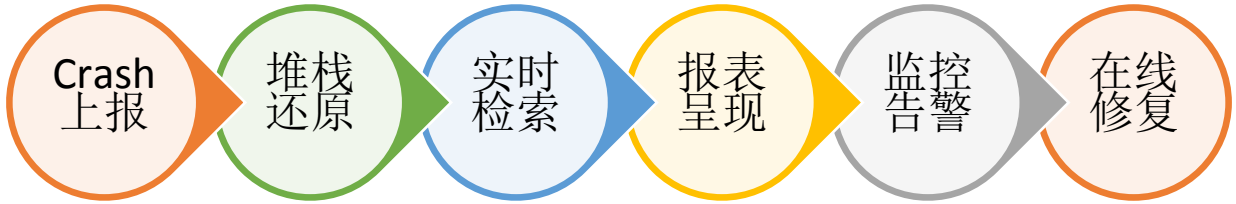
通道切换 | 配置下发

热更新 | 控制指令

基础服务质量监控

SDK 质量 | 网络服务质量

终端质量监控



运营平台
- 监控告警
- 复现告警
- 报告订阅
- 符号表管理
- 云控系统

Android
- Native
- Java

iOS
- Object-c
- Swift

细分领域
- Cocos2dX
- Unity
- LUR
- JS

数据服务
- crash 质量分析报告
- 质量波动预警

Access 统一接入层

数据网关

云控
- 云端指令
- 配置下发
- 热更新
- 路由规则
  - App
  - 设备ID

TDBank 消息中间件

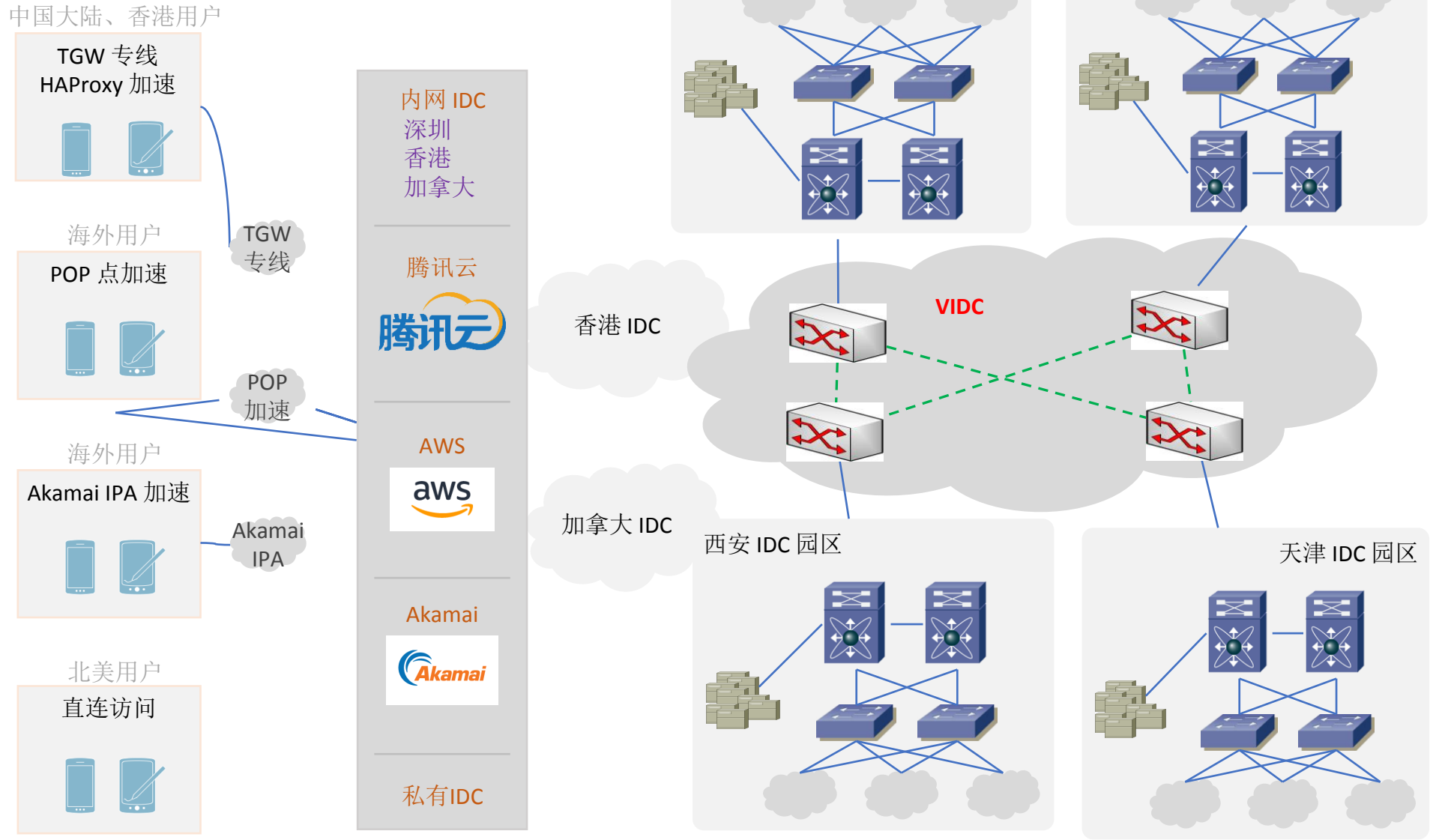
批量处理
- Spark Streaming
- MR(离线备用方案)

HDFS | Redis | PGX

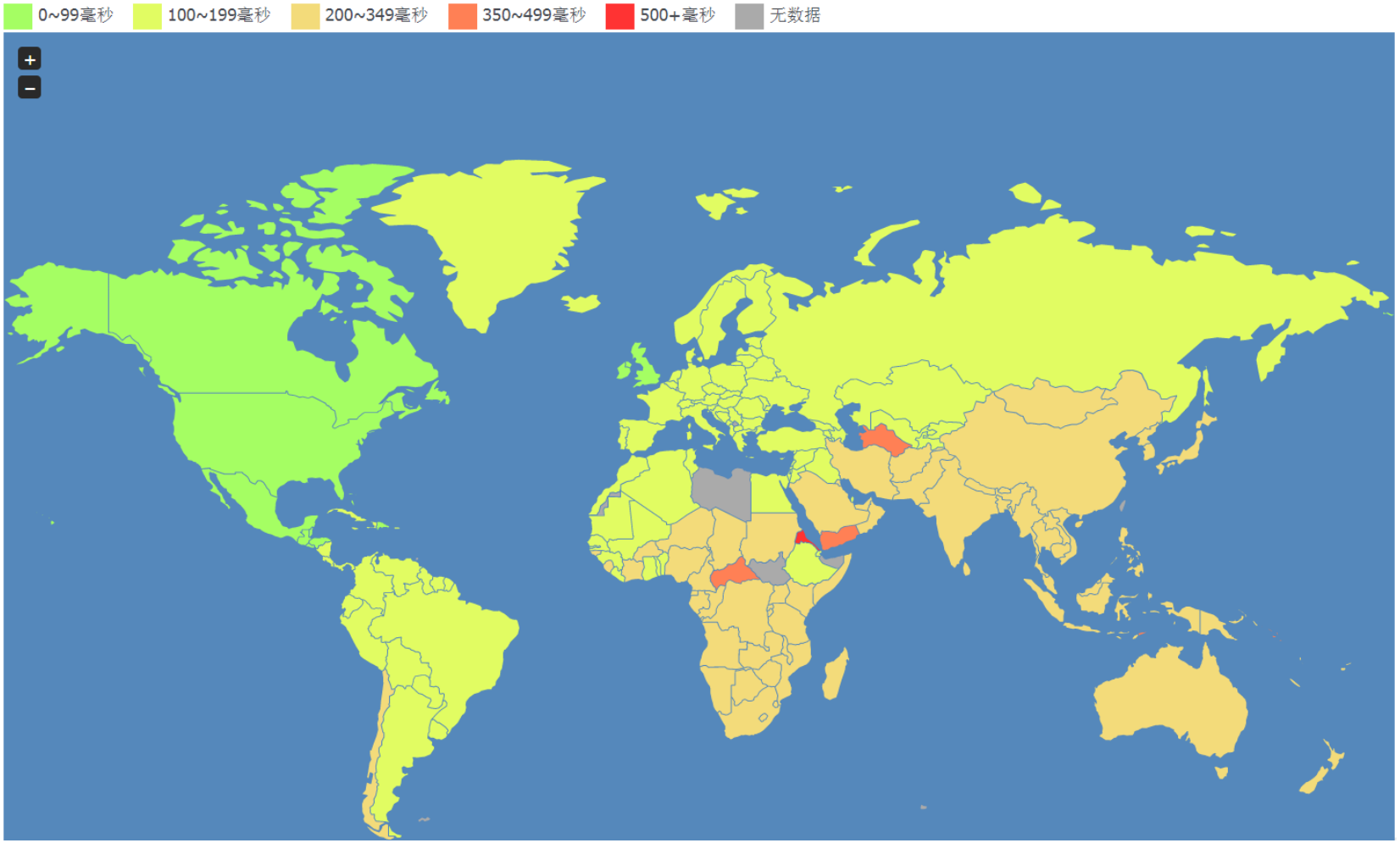
实时处理
- 堆栈还原
  - iOS系统
  - 符号表
- ElasticSearch
  - 全文检索
  - 聚合统计
- 实时计算
  - ESP
  - ESA
  - ESC

数据挖掘
- crash 聚类
- 智能专家
- 众包生态
- 质量预警

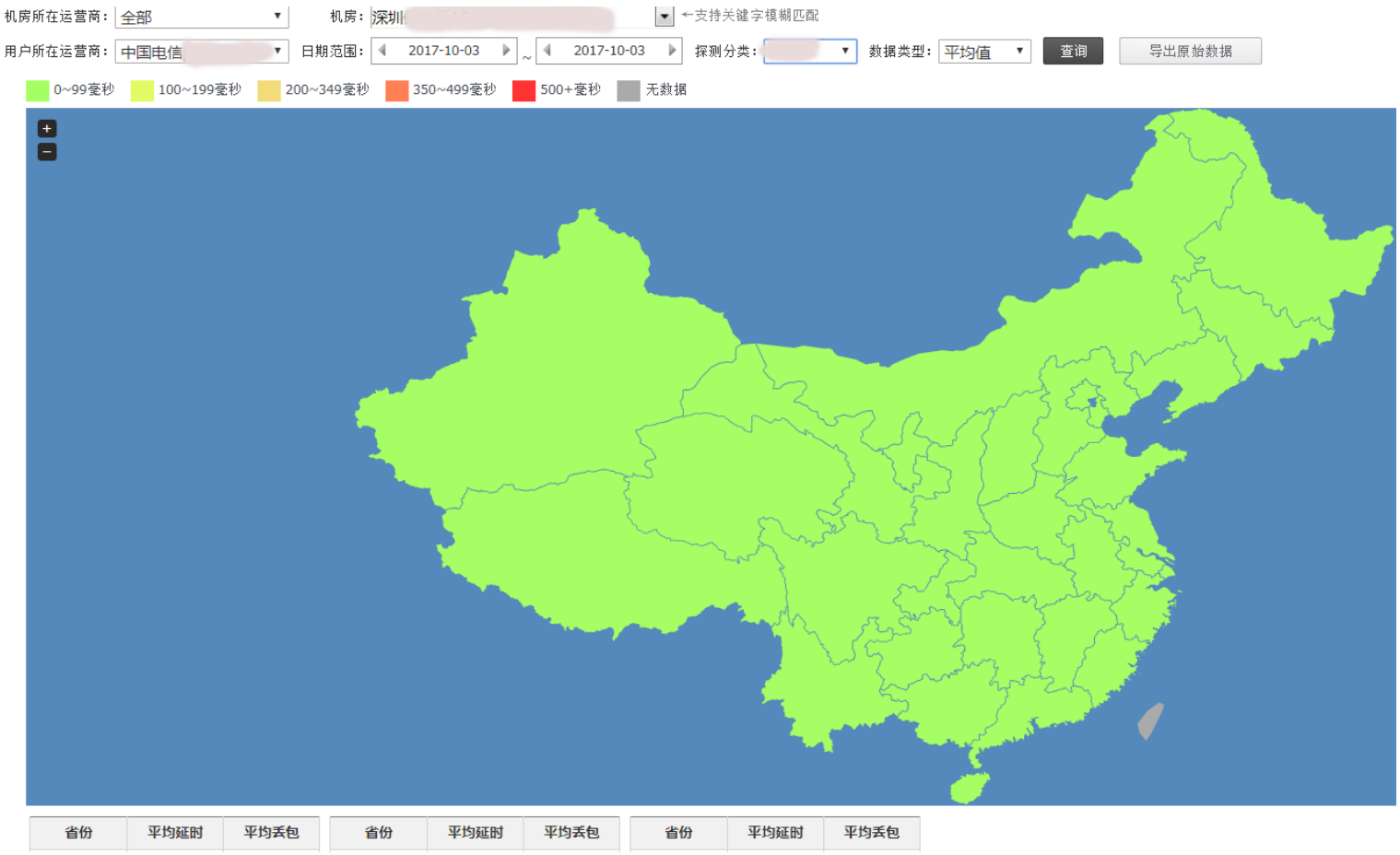
网络基础设施


中国大陆、香港用户
TGW 专线
HAProxy 加速
海外用户
POP 点加速
海外用户
Akamai IPA 加速
北美用户
直连访问
TGW 专线
POP 加速
Akamai IPA
内网 IDC
深圳
香港
加拿大
腾讯云
腾讯云
AWS
aws
Akamai
Akamai
私有IDC
深圳 IDC 园区
上海 IDC 园区
香港 IDC
VIDC
加拿大 IDC
西安 IDC 园区
天津 IDC 园区

全球网络服务质量监控 – 海外用户视角

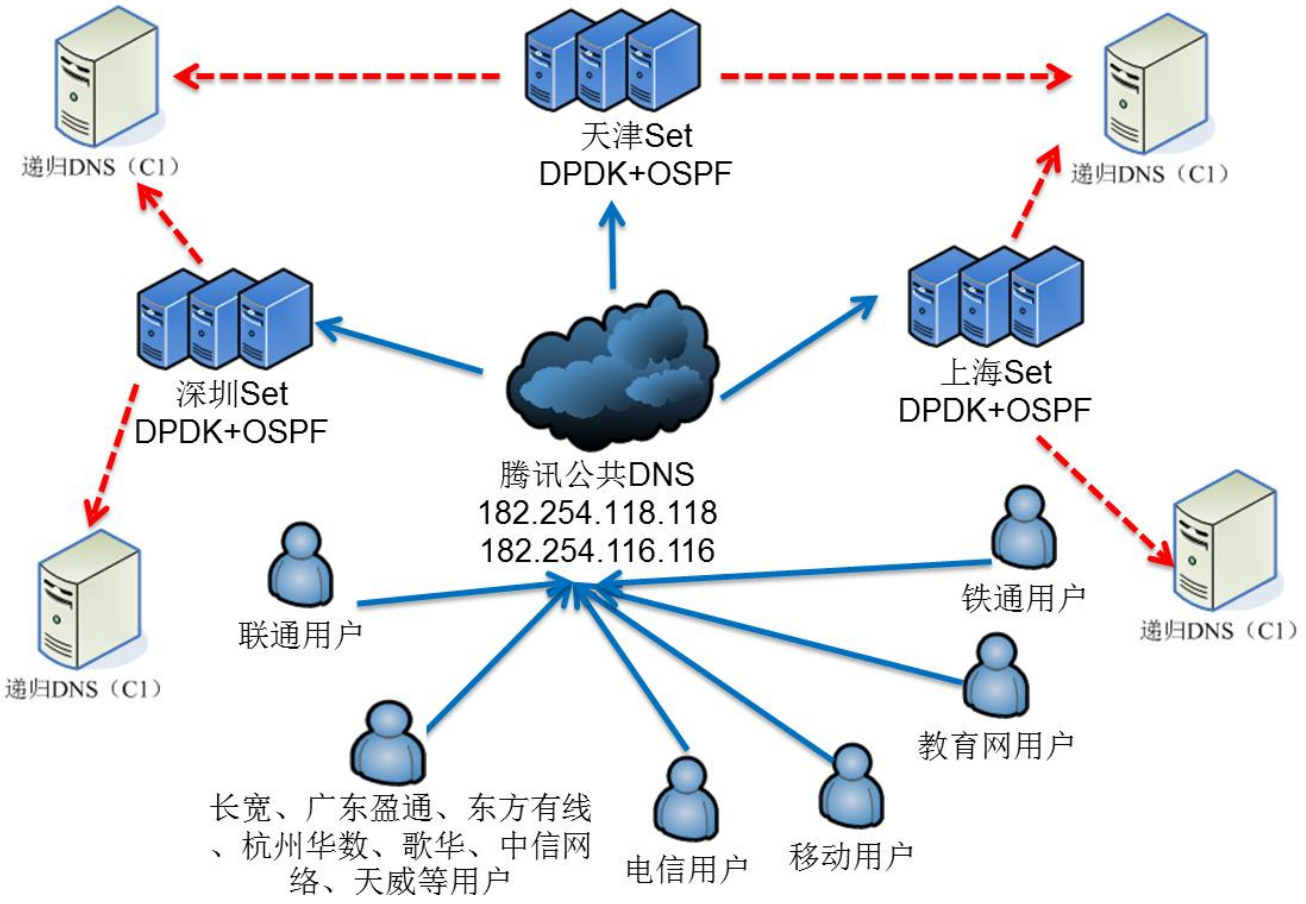


国内网络服务质量监控 – 国内用户视角



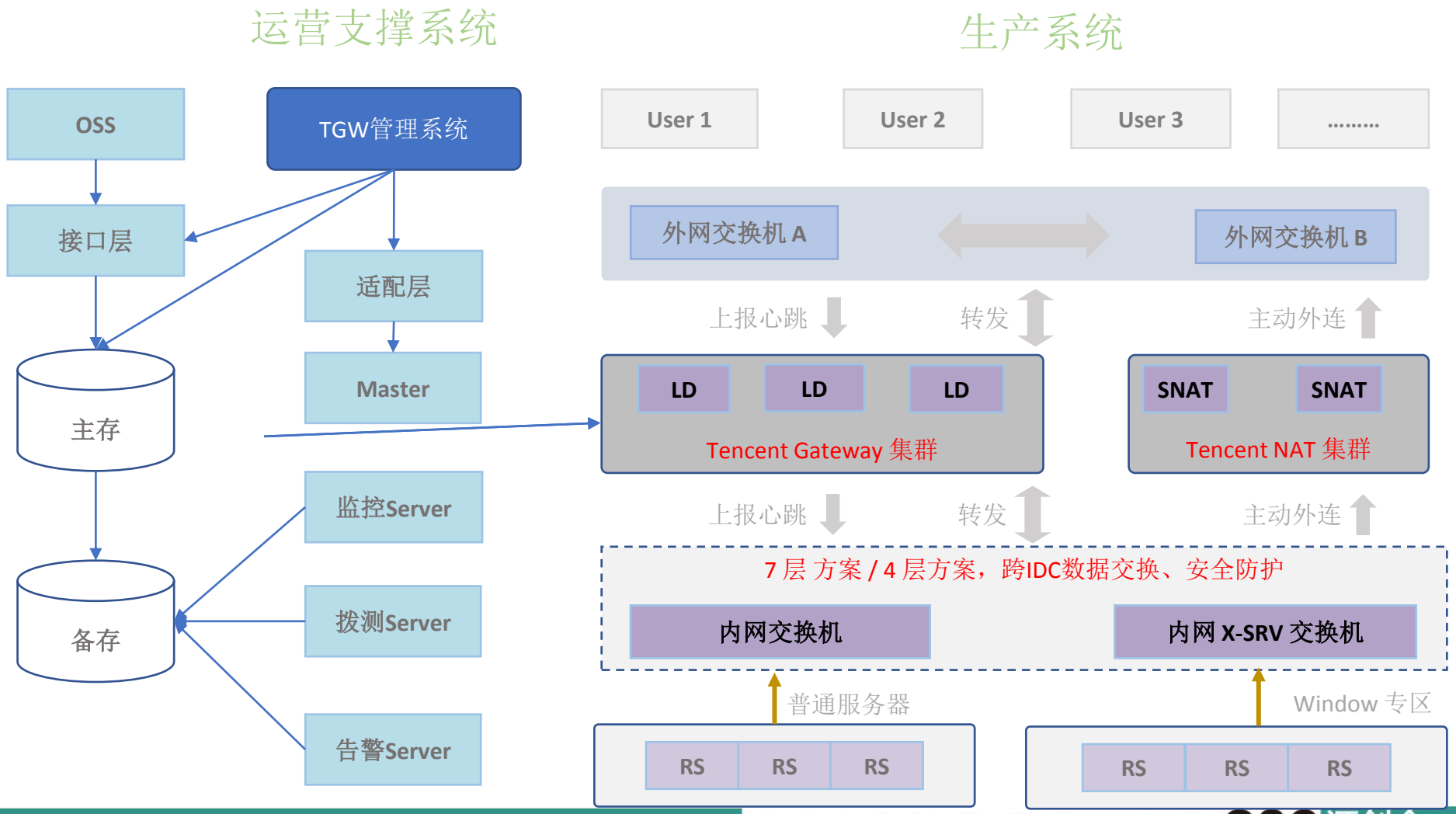
自建DNS服务

一个IP、多地容灾，OSPF集群IDC独立部署，流量秒级无感知切换。



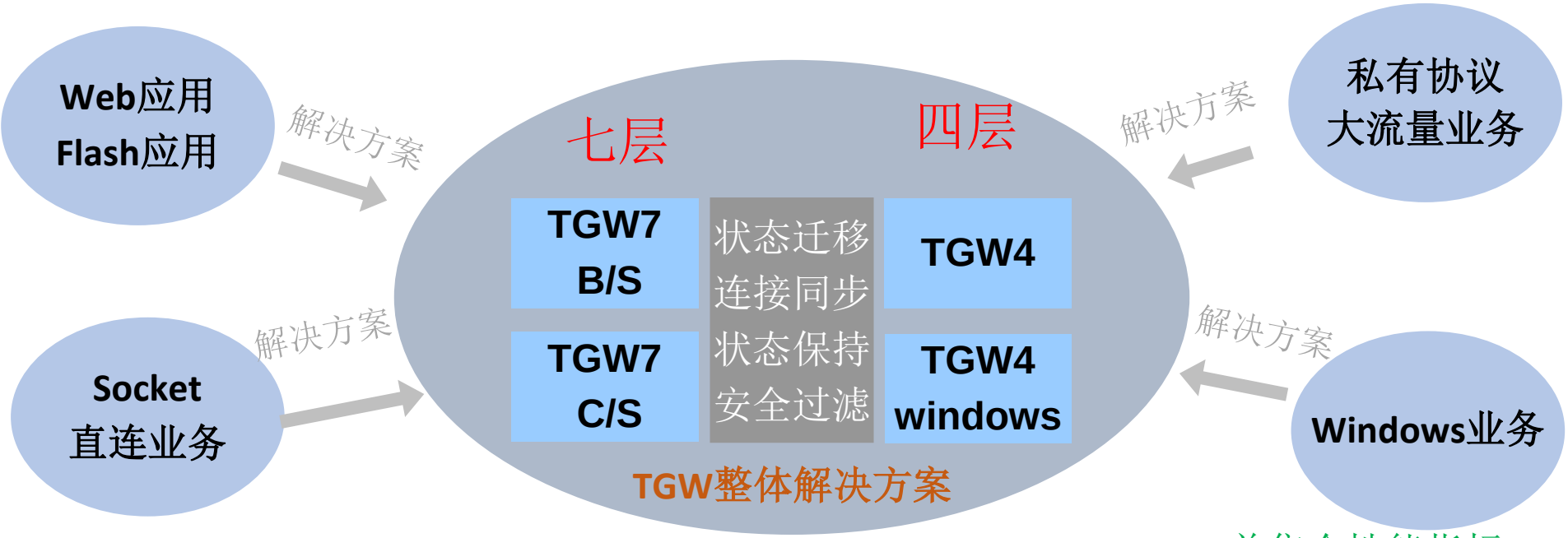
统一接入网关

统一接入、多网合一；负载均衡、平滑扩容；网络隔离防护、IP 收敛。



四层、七层接入方案

四种方案：七层B/S方案， 七层C/S方案，四层方案，四层windows方案。



TGW SET模型	LD数	最大容量	最大包量	接入VIP数	最多接入域名个数	最多接入RS个数
TGW7-1G	4	2G	300w	4	800	800
TGW7-10G	4	10G	300w	20	4000	4000
TGW4-10G	4	16G	500w	N/A	N/A	N/A
TGW4-NAT	4	1G	150w	N/A	N/A	N/A

单集合性能指标

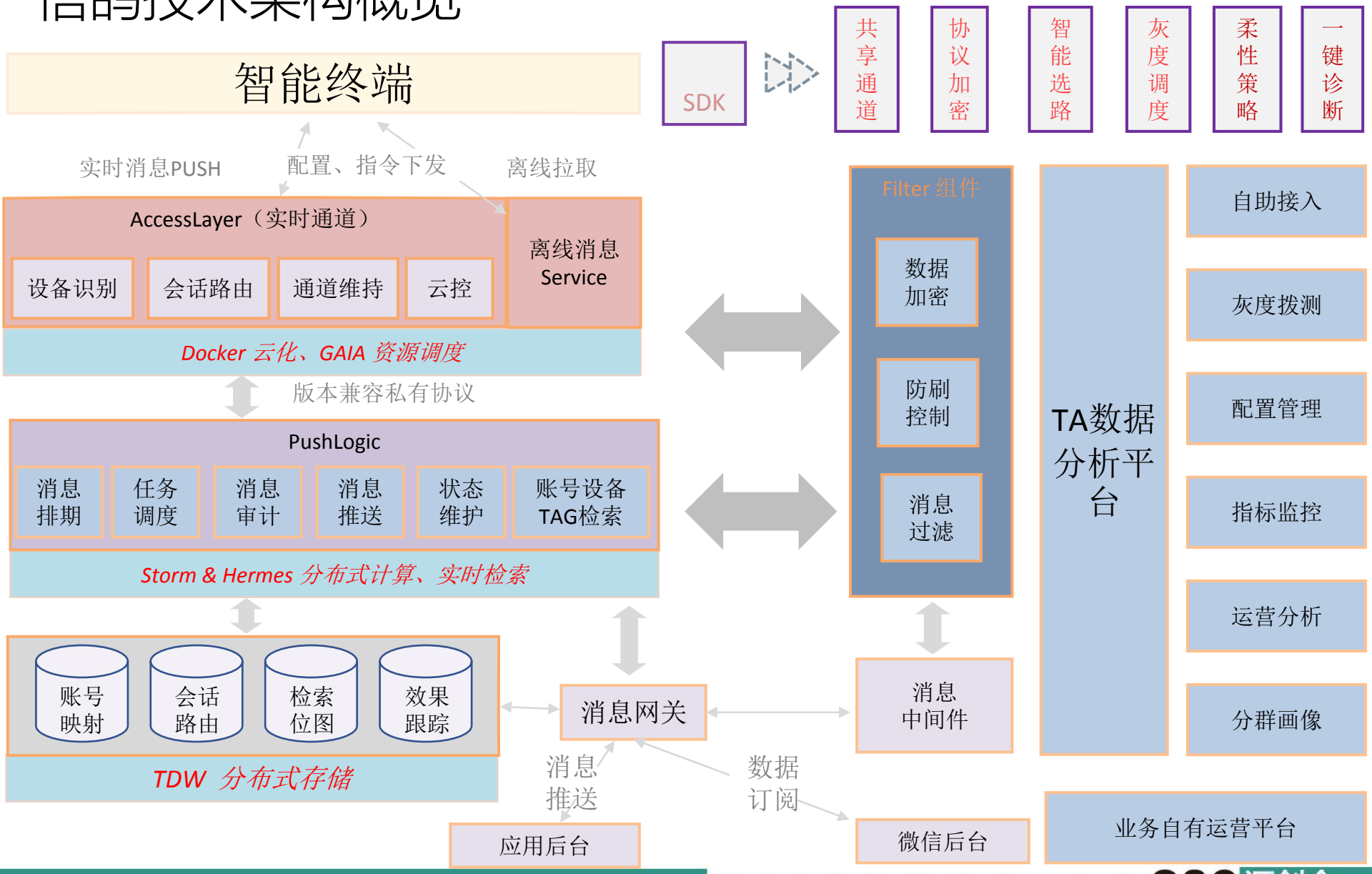
延时：< X ms

并发连接数：XXX万

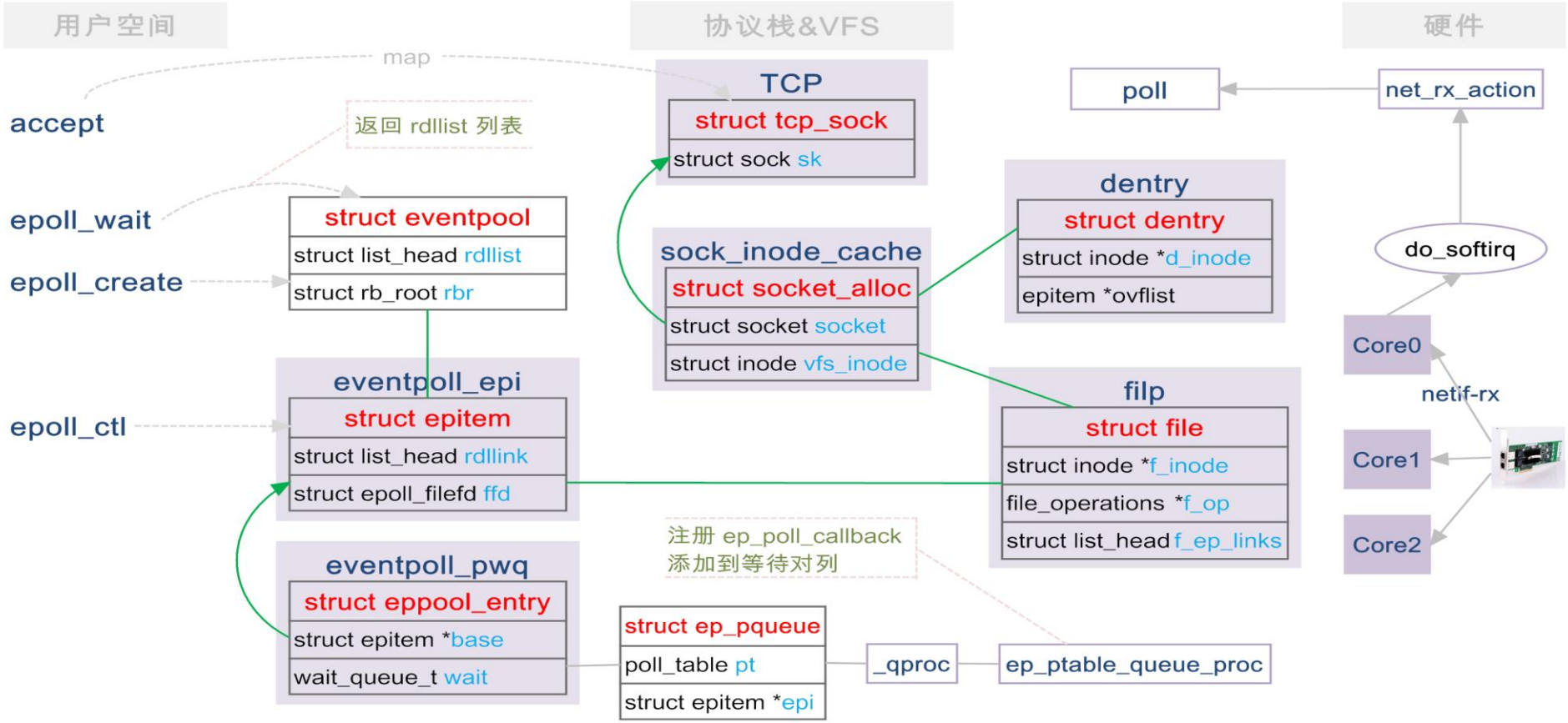
包量：XXX万 PPS

带宽：XX Gps

信鸽技术架构概览


智能终端
SDK
共享通道
协议加密
智能选路
灰度调度
柔性策略
一键诊断
实时消息PUSH
配置、指令下发
离线拉取
AccessLayer（实时通道）
设备识别
会话路由
通道维持
云控
离线消息 Service
Docker 云化、GAIA 资源调度
版本兼容私有协议
PushLogic
消息排期
任务调度
消息审计
消息推送
状态维护
账号设备TAG检索
Storm & Hermes 分布式计算、实时检索
账号映射
会话路由
检索位图
效果跟踪
TDW 分布式存储
消息网关
消息推送
数据订阅
应用后台
Filter 组件
数据加密
防刷控制
消息过滤
消息中间件
微信后台
TA数据分析平台
自助接入
灰度拨测
配置管理
指标监控
运营分析
分群画像
业务自有运营平台

关键的6个 kernel memory cache



单节点接入百万级长连接

C1 机型 8G 内存，支持 200W+ 长连接

```
 Active / Total Objects (% used)    : 15078299 / 15185017 (99.3%)
 Active / Total Slabs (% used)      : 1437021 / 1437093 (100.0%)
 Active / Total Caches (% used)     : 90 / 215 (41.9%)
 Active / Total Size (% used)       : 6638525.50K / 6651639.44K (99.8%)
 Minimum / Average / Maximum Object : 0.02K / 0.44K / 4096.00K

  OBJS ACTIVE  USE OBJ SIZE  SLABS OBJ/SLAB CACHE SIZE NAME
3492460 3491904  99%    0.19K 174623       20    698492K dentry
2001840 2001624  99%    0.19K 100092       20    400368K filp
2000273 1999995  99%    0.07K  37741       53    150964K eventpoll_pwq
2000100 1999995  99%    0.12K  66670       30    266680K eventpoll_epi
2000095 2000084  99%    0.69K 400019        5   1600076K sock_inode_cache
2000020 1999996  99%    1.44K 400004        5   3200032K TCP
1490376 1490039  99%    0.63K 248396        6    993584K proc_inode_cache
```

- TCP（1.6K）：关联了socket、sock、file
- sock_inode_cache（0.8K）：socket_alloc，包含 socket 和 inode 节点
- eventpoll_epi（0.13K）：epitem，
- eventpoll_pwq（0.07K）：epoll_entry 负责关联fd等待事件和epitem
- filp（0.2K）：文件指针
- dentry（0.2K）：dentry

创建一个sock，到Epoll监听，内核消耗 **3K**，其中结构体占 2.7K，对齐或 slab 浪费 0.3K

并行检索、推送能力

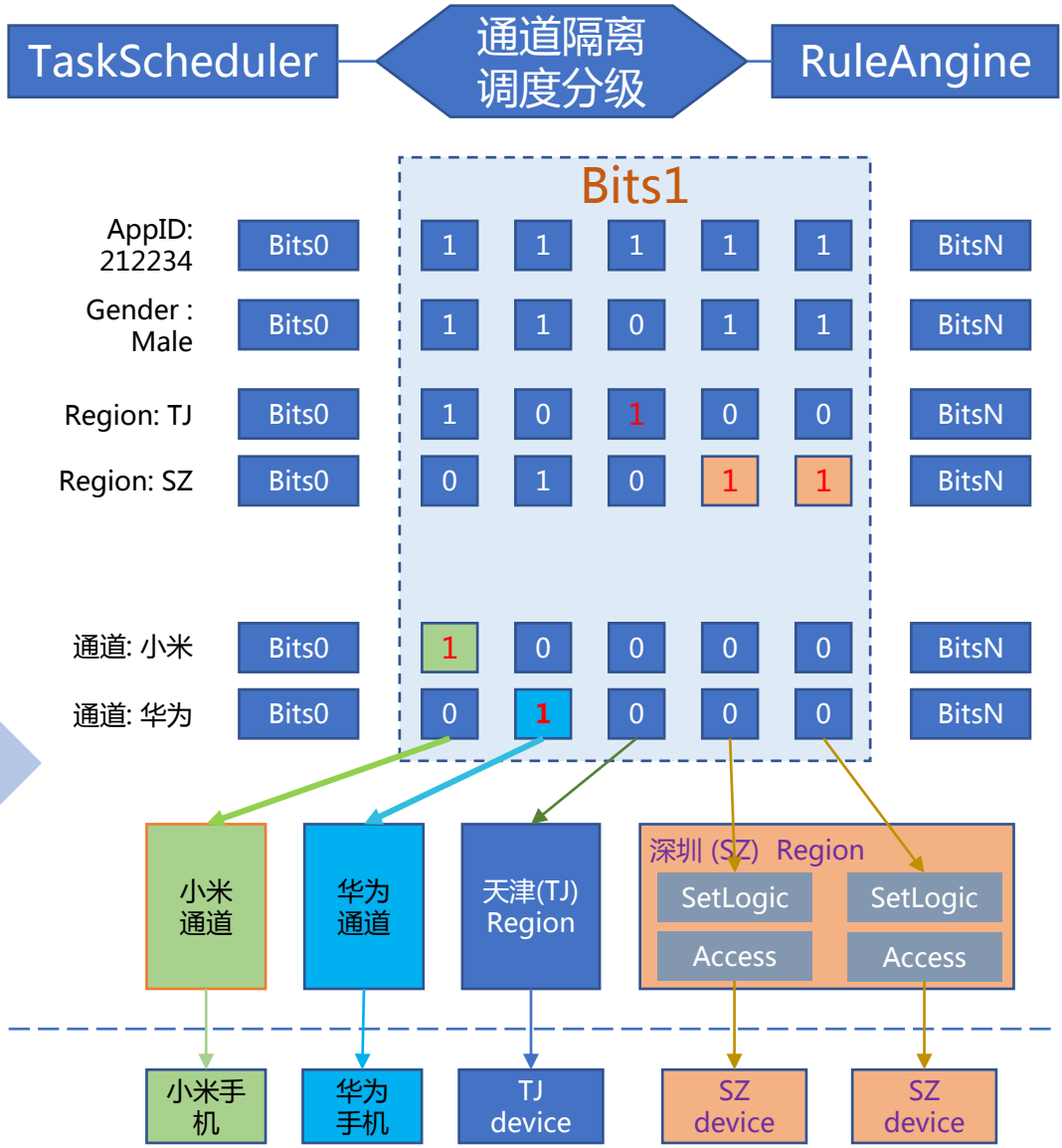
逻辑表达转换为析取范式

$$(A \cup B) \cap C \Rightarrow (A \cap C) \cup (B \cap C)$$

工程问题映射成数学问题

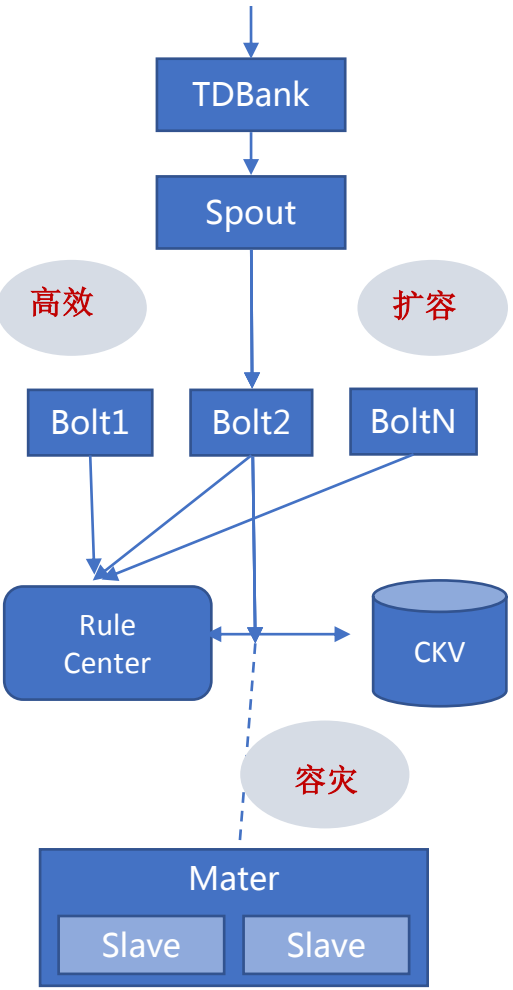
- 查询：查询需推送设备
- 排期：任务、数据分割准备
- 调度：申请资源、调度执行
- 路由：选择通道推送
- 推送：消费队列、触达用户

数据传输、计算、查询量大。
逐层下漏，层之间容易形成瓶颈。



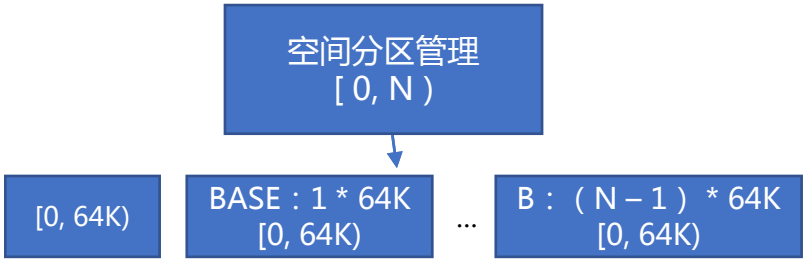
推送效果实时统计

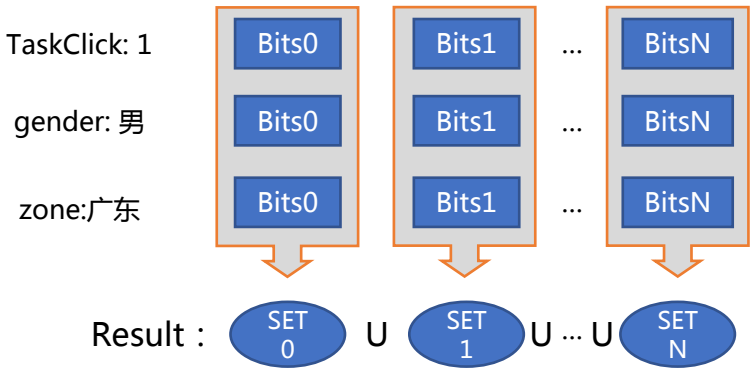
文本数值化



高效数据结构

RoaringBitmap DataPartition

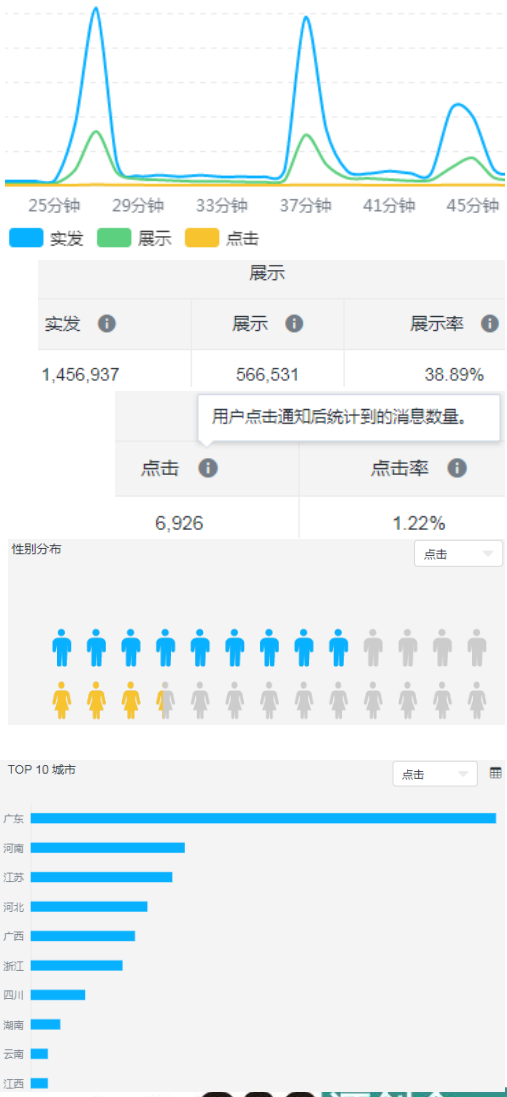




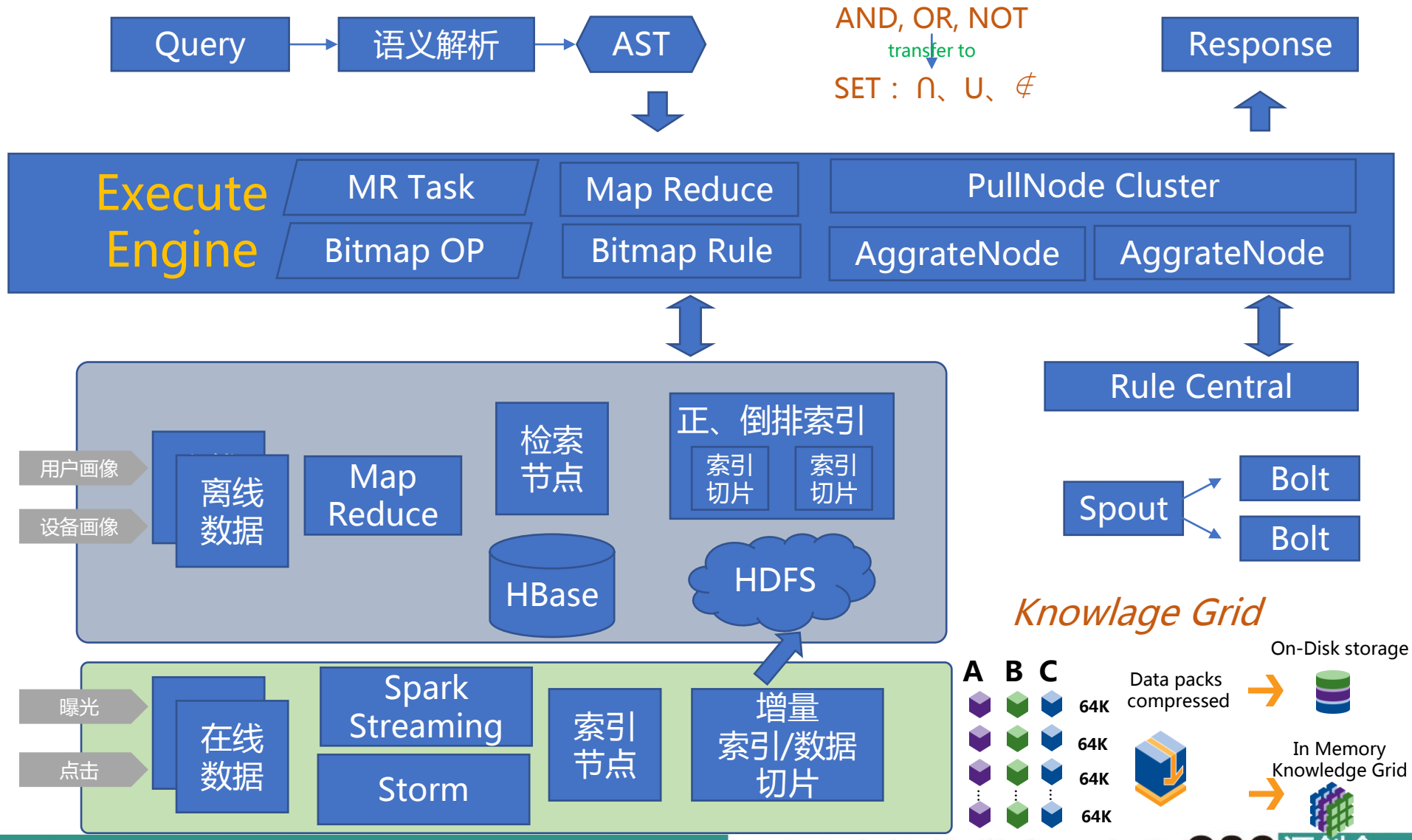
广东、男性用户点击

拿设备标识举例：40 字节，**320 bit** 到 **1 bit** 的压缩

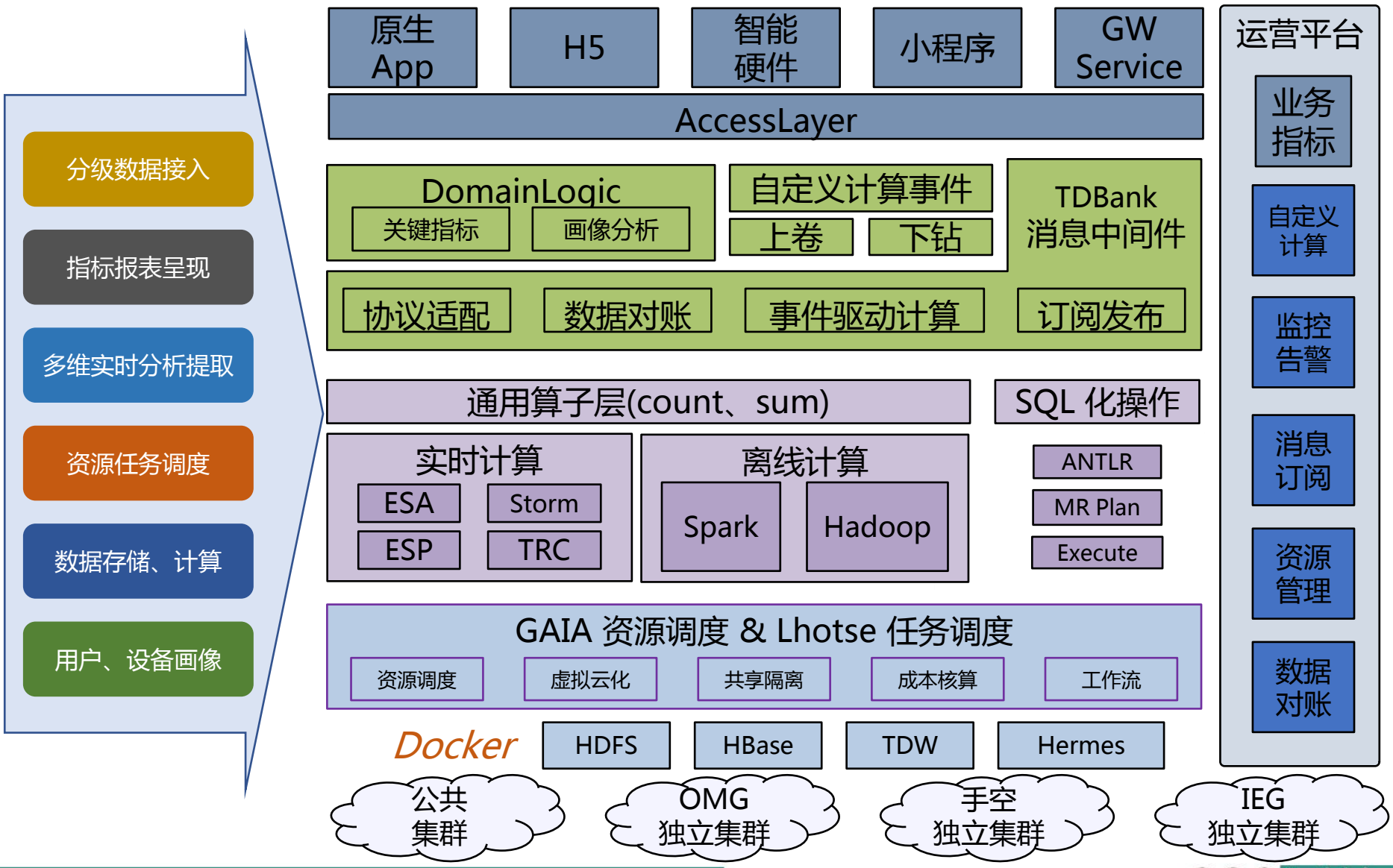
实时效果统计



海量数据，实时多维分析


Query
语义解析
AST
AND, OR, NOT
transfer to
SET : ∩、U、∉
Response
Execute Engine
MR Task
Bitmap OP
Map Reduce
Bitmap Rule
PullNode Cluster
AggrateNode
AggrateNode
Rule Central
用户画像
设备画像
离线数据
Map Reduce
检索节点
正、倒排索引
索引切片
索引切片
HBase
HDFS
Spout
Bolt
Bolt
Knowlage Grid
曝光
点击
在线数据
Spark Streaming
Storm
索引节点
增量索引/数据切片
A B C
64K
64K
64K
64K
Data packs compressed
On-Disk storage
In Memory Knowledge Grid

数据分析平台


分级数据接入
指标报表呈现
多维实时分析提取
资源任务调度
数据存储、计算
用户、设备画像
原生App
H5
智能硬件
小程序
GW Service
AccessLayer
DomainLogic
关键指标
画像分析
自定义计算事件
上卷
下钻
TDBank 消息中间件
协议适配
数据对账
事件驱动计算
订阅发布
通用算子层(count、sum)
SQL 化操作
实时计算
ESA
Storm
ESP
TRC
离线计算
Spark
Hadoop
ANTLR
MR Plan
Execute
GAIA 资源调度 & Lhotse 任务调度
资源调度
虚拟云化
共享隔离
成本核算
工作流
Docker
HDFS
HBase
TDW
Hermes
公共集群
OMG 独立集群
手空 独立集群
IEG 独立集群
运营平台
业务指标
自定义计算
监控告警
消息订阅
资源管理
数据对账

DevOps

MySQL

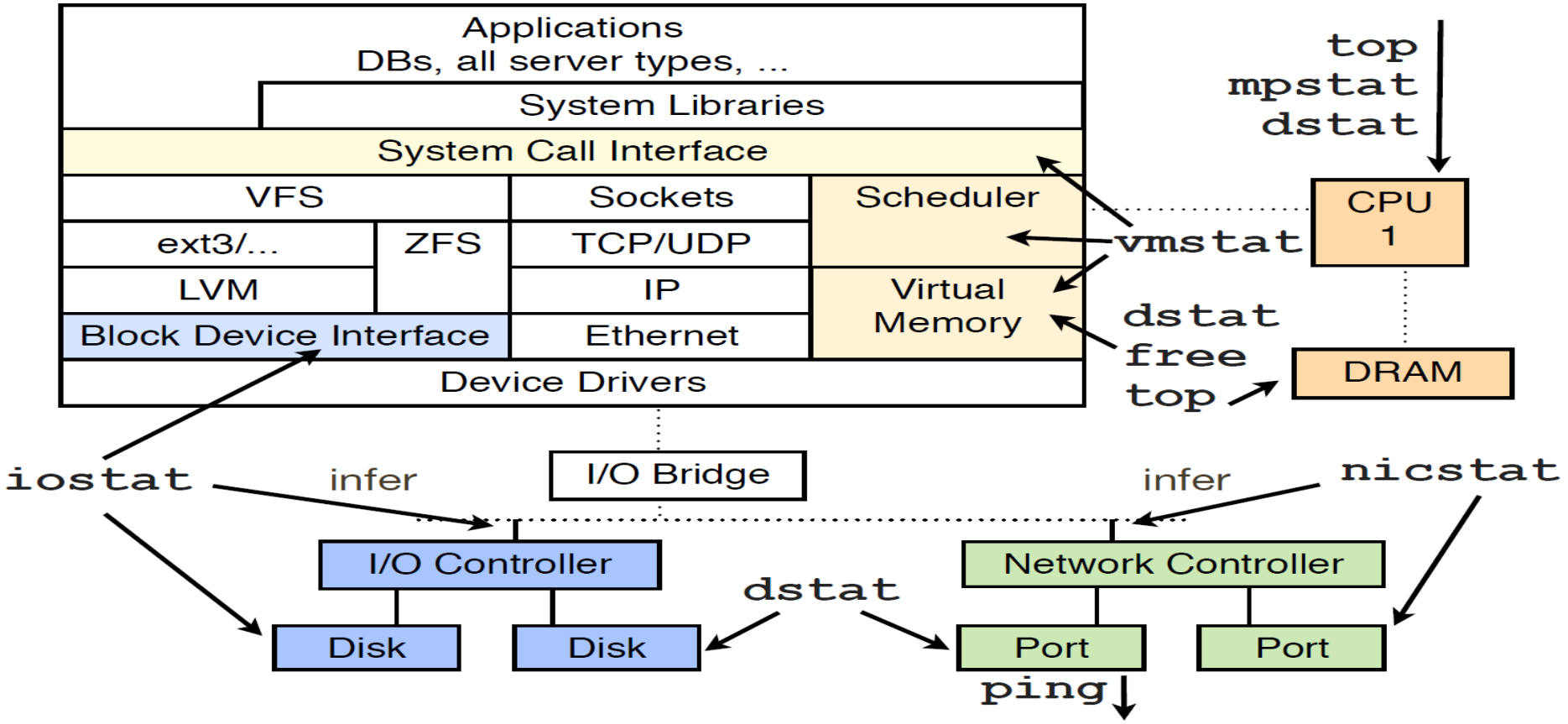
配置中心

L5

服务发现负载均衡

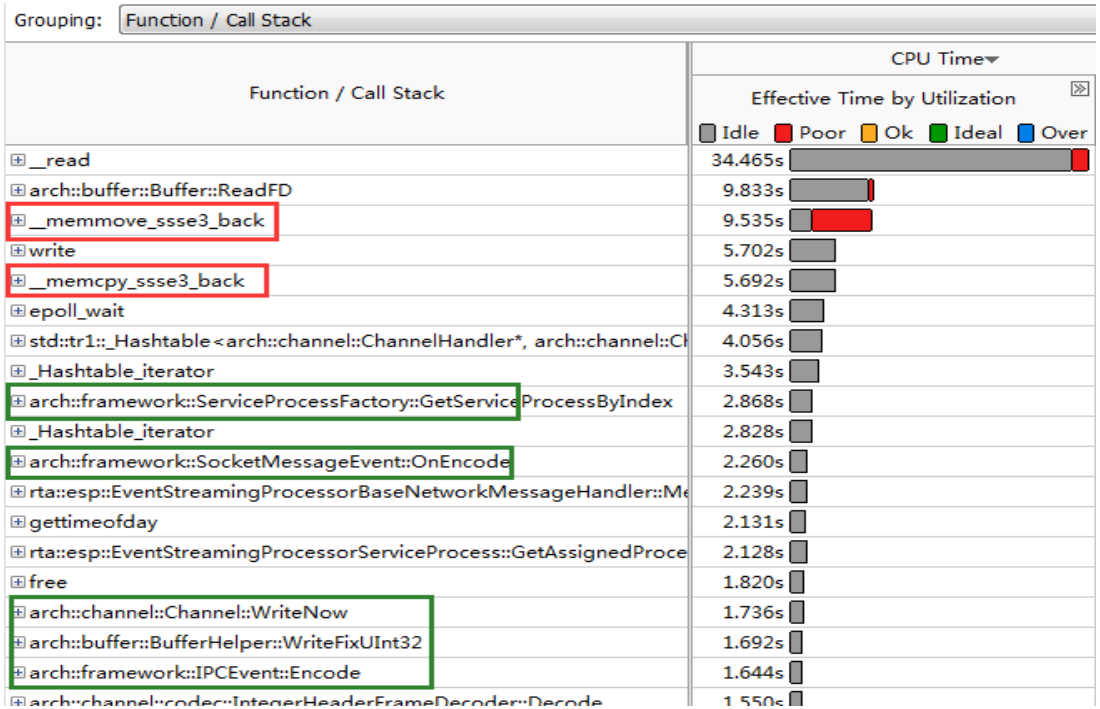
GAIA STACK

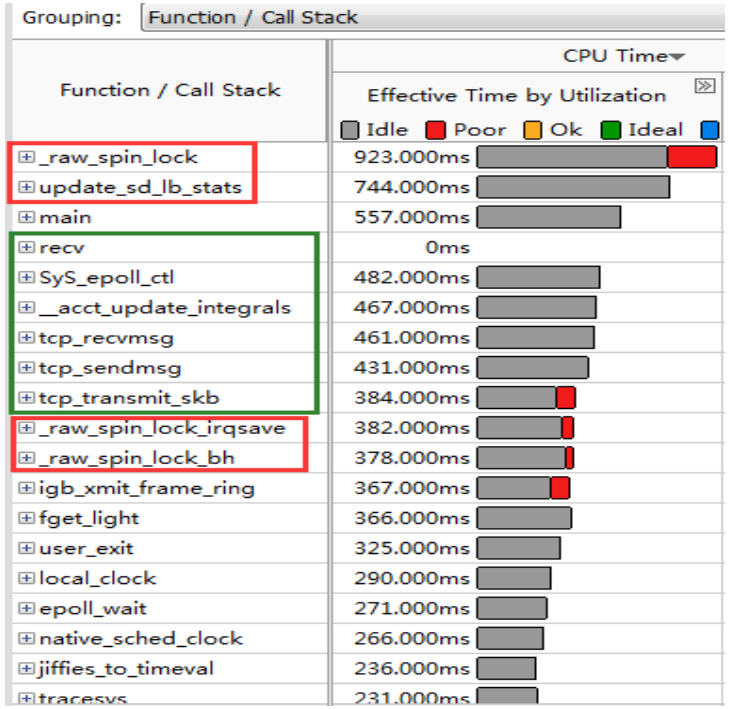
硬件、协议栈、内核、用户态不同层次评测



摘自：Systems Performance Enterprise and the Cloud

单机性能挖掘





较好的网络框架，耗时应该主要集中在网络的处理上，而非框架本身。

基于 **Linux协议栈** 网络服务，大部分性能消耗在 **spin_lock & 网络协议栈** 上。

单机性能挖掘 — 评测

amplxe-cl -collect advanced-hotspots -result-dir rs01 ./access_proxy

Function / Call Stack	CPU Time: Effective Time by Utilization (Idle Poor Ok Ideal Over)	CPI Rat.	CPU Fre... Ratio	Module	Function (Full)
_raw_spin_lock	722.000ms	2.6..	0.853	vmlinux	_raw_spin_lock
SyS_epoll_ctl	379.000ms	3.7..	0.592	vmlinux	SyS_epoll_ctl
update_sd_lb_stats	235.000ms	1.3..	0.682	vmlinux	update_sd_lb_stats
__acct_update_integrals	221.000ms	0.7..	0.566	vmlinux	__acct_update_integ...
_raw_spin_lock_bh	203.000ms	1.9..	0.582	vmlinux	_raw_spin_lock_bh
user_exit	191.000ms	1.0..	0.514	vmlinux	user_exit
kmem_cache_alloc	170.000ms	1.8..	0.448	vmlinux	kmem_cache_alloc
main	170.000ms	3.2..	0.607	access_pr...	main
_raw_spin_lock_irqsave	167.000ms	1.4..	0.743	vmlinux	_raw_spin_lock_irqs...
epoll_wait	163.000ms	43..	0.528	lib...	
jiffies_to_timeval	155.000ms	0.8..	0.452	vm	
native_sched_clock	127.000ms	1.4..	0.583	vm	
local_clock	122.000ms	0.5..	0.591	vm	
ip_queue_xmit	121.000ms	3.0..	1.100	vm	
tcp_transmit_skb	117.000ms	1.0..	1.138	vm	
MD::BinUtil::crc32	116.000ms	0.9..	0.699	ac	
igb_xmit_frame_ring	111.000ms	0.6..	1.001	igb	
sched_clock_cpu	106.000ms	0.6..	0.482	vm	
tcp_recvmsg	103.000ms	1.6..	0.573	vm	

perf top –p pid

```
Samples: 377K of event 'cycles', Event count (approx.): 797075664
  3.10%  [kernel]              [k] update_sd_lb_stats
  2.46%  [kernel]              [k] _raw_spin_lock
  2.45%  access_proxy          [.] main
  2.20%  [kernel]              [k] fget_light
  2.03%  [kernel]              [k] sys_epoll_ctl
  2.01%  [kernel]              [k] tcp_sendmsg
  1.84%  [kernel]              [k] tcp_recvmsg
  1.70%  [kernel]              [k] _raw_spin_lock_bh
  1.55%  [kernel]              [k] tcp_transmit_skb
  1.41%  [igb]                 [k] igb_xmit_frame_ring
  1.20%  libc-2.17.so          [.] __memcpy_ssse3_back
  1.01%  [kernel]              [k] _raw_spin_lock_irqsave
  0.93%  [kernel]              [k] ep_send_events_proc
  0.93%  [ip_tables]           [k] ipt_do_table
  0.92%  [kernel]              [k] tcp_write_xmit
  0.90%  access_proxy          [.] handleSocketMessage(InternetMessa
  0.89%  [kernel]              [k] find_next_bit
  0.89%  [kernel]              [k] system_call
  0.87%  [kernel]              [k] __schedule
  0.86%  [kernel]              [k] update_blocked_averages
  0.79%  [kernel]              [k] tcp_poll
  0.76%  [kernel]              [k] tcp_cleanup_rbuf
```

优化目标：

- ✓ 优化函数执行效率
- ✓ 消除内核协议栈消耗
- ✓ 挖掘硬件潜能

协议栈改造


```
--------------------------------------------------------------------------------
Interfaces:
name: eth0, ifindex: 0, hwaddr: 54:89:98:72:5B:1E, ipaddr: 10.125.225.3, netmask: 255.255.
Number of NIC queues: 4
--------------------------------------------------------------------------------
Loading routing configurations from : config/route.conf
fopen: No such file or directory
Skip loading static routing table
Routes:
Destination: 10.125.225.0/32, Mask: 255.255.255.192, Masked: 10.125.225.0, Route: ifdx-0
--------------------------------------------------------------------------------
Loading ARP table from : config/arp.conf
fopen: No such file or directory
Skip loading static ARP table
ARP Table:
(blank)
EAL: Master lcore 0 is ready (tid=ee0d18c0;cpuset=[0])
EAL: lcore 1 is ready (tid=ebbe7700;cpuset=[1])
EAL: lcore 2 is ready (tid=eb3e6700;cpuset=[2])
EAL: lcore 3 is ready (tid=eabe5700;cpuset=[3])
EAL: PCI device 0000:01:00.0 on NUMA socket -1
EAL:   probe driver: 8086:10c9 rte_igb_pmd
EAL:   PCI memory mapped at 0x7fc2ec800000
EAL:   PCI memory mapped at 0x7fc2ec820000
EAL:   PCI memory mapped at 0x7fc2ec840000
PMD: eth_igb_dev_init(): port_id 0 vendorID=0x8086 deviceID=0x10c9
EAL: PCI device 0000:01:00.1 on NUMA socket -1
EAL:   probe driver: 8086:10c9 rte_igb_pmd
EAL:   Not managed by a supported kernel driver, skipped
Interactive-mode selected
Configuring Port 0 (socket 0)
PMD: eth_igb_tx_queue_setup(): To improve 1G driver performance, consider setting the TX W
PMD: eth_igb_tx_queue_setup(): sw_ring=0x7fc1ec5a7e40 hw_ring=0x7fc1ec5a9e80 dma_addr=0x22
PMD: eth_igb_rx_queue_setup(): sw_ring=0x7fc1ec5978c0 hw_ring=0x7fc1ec597d00 dma_addr=0x22
PMD: eth_igb_start(): <<
Port 0: 54:89:98:72:5B:1E
Checking link statuses...
Port 0 Link Up - speed 1000 Mbps - full-duplex
Done
```

协议栈改造

A： Kernel Socket 函数调用

```
int iSock = socket(PF_INET, SOCK_STREAM, 0);
```

B： mTCP 的函数函数调用

```
int iSock = mtcp_socket(g_pMctx, PF_INET, SOCK_STREAM, 0);
```

C： 业务无感知的 mTCP 协议栈改造

```
typedef int (mTcpHookSock*)(int socket_family, int socket_type, int protocol);
int realSocketFunc(int socket_family, int socket_type, int protocol) {
    return mtcp_socket(g_pMctx, socket_family, socket_type, protocol);
}

typedef struct {
    mTcpHookSock      hookSockFunc;
    mTcpHookAccept    hookAcceptFunc;
} stMTCPHooFuncTable;

g_syncHookTable.hookSockFunc = dlsym(RTLD_NEXT, "socket");
```

拦截系统调用

指向 mtcp_socket WRAP 函数

硬件性能挖掘

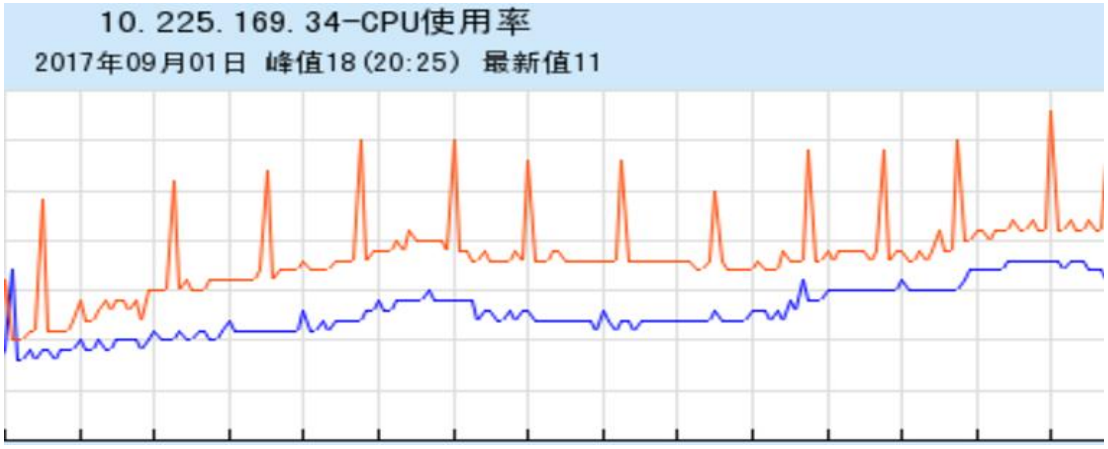
Table Lookup → SSE 4.2 _mm_crc32_u32 → SSE 4.2 Parallel crc32q

local_clock	122.000ms	308,660,000
ip_queue_xmit	121.000ms	111,320,000
tcp_transmit_skb	117.000ms	331,430,000
MD::BinUtil::crc32	116.000ms	212,520,000
igb_xmit_frame_ring	111.000ms	427,570,000
sched_clock_cpu	106.000ms	199,870,000
tcp_recvmsg	103.000ms	88,550,000

__errno_location	0ms	5,060,000
update_cfs_shares	20.000ms	15,180,000
retransmits_timed_out	20.000ms	32,890,000
MD::BinUtil::crc32csse	20.000ms	43,010,000
ipv4_dst_check	20.000ms	43,010,000
ip_output	20.000ms	50,600,000
inode_init_once	20.000ms	22,770,000

执行 35W 次计算结果（SSE Parallel crc32q 快接近 6 倍）

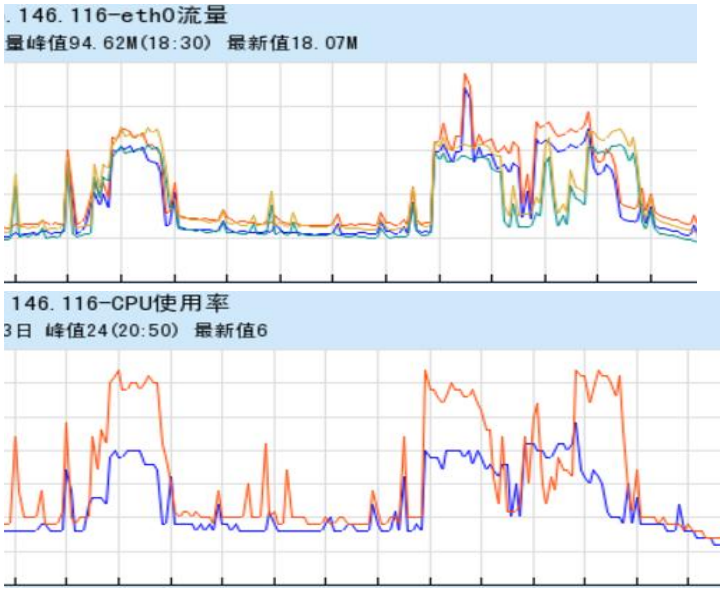
__memcpy_ssse3_back & __mempcpy_avx_unaligned 、 AES_cfb8_encrypt & evp_encrypt (M10 机型)




```
typedef uint32_t (CRC32FUNC)(const char* pBuf, int iLen);

typedef struct {
    CRC32FUNC       crc32;
} stOptimizeLayer;

uint32_t eax, ebx, ecx, edx;

eax = ebx = ecx = edx = 0;
__get_cpuid(1, &eax, &ebx, &ecx, &edx);
if ( ecx & bit_SSE4_2 ) {
    opts.crc32      =       sse42_crc32;
} else {
    opts.crc32      =       default_crc32;
}

uint32_t crc32(const char* pBuf, int iLen)
            __attribute__( (ifunc("resolve_crc32")) );
```

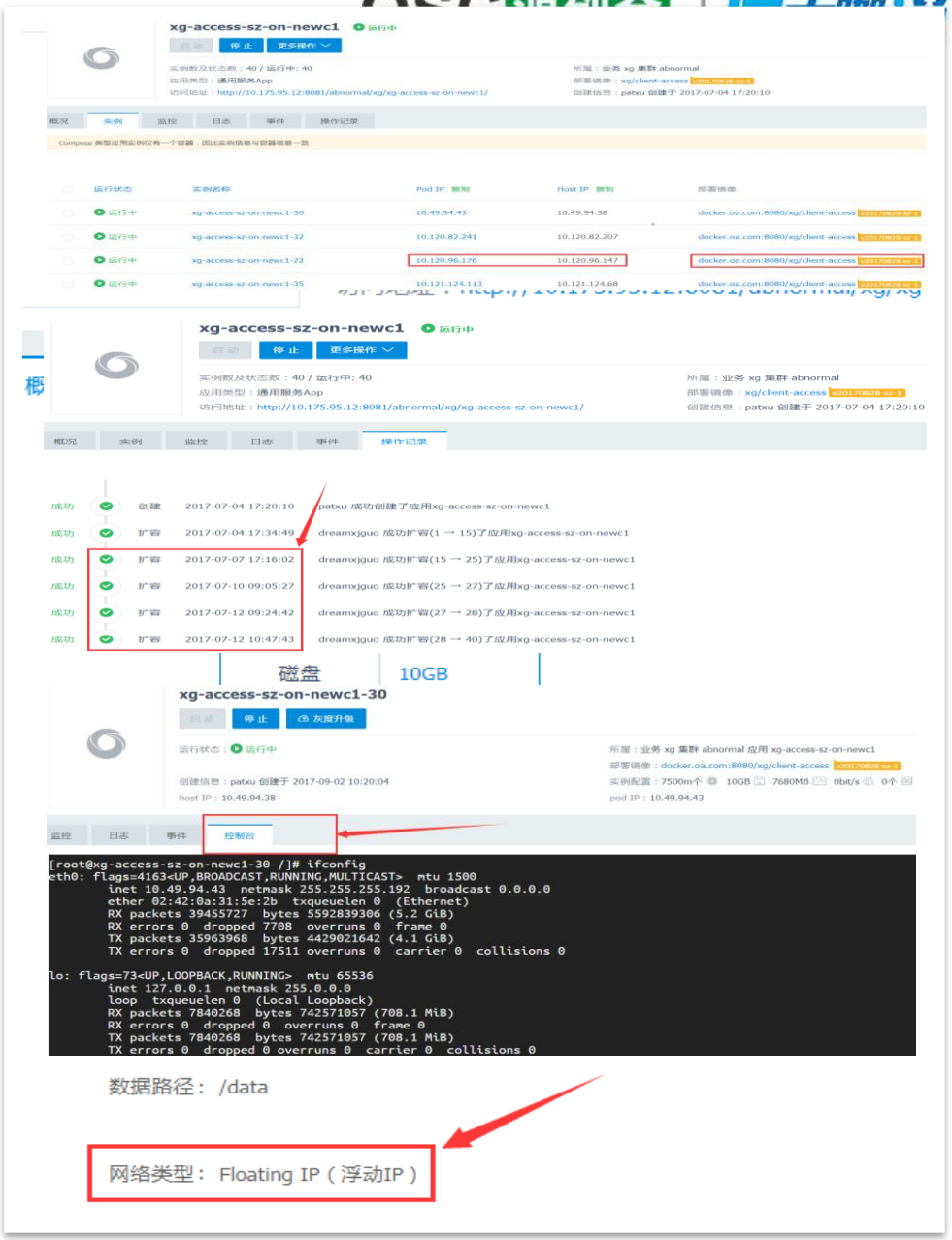
硬件虚拟化提升效能

docker & SR-IOV，高峰 CPU 降低约 38.3%



信鸽、MTA Docker 云化部署

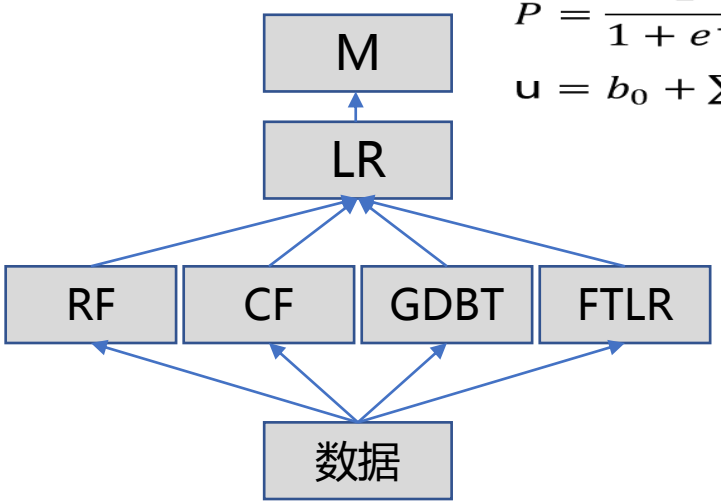
1. 混合部署、资源隔离
2. AutoScaling，自动加入 TGW、L5
3. Floating IP
4. 网络层启用 SR-IOV 特性



信鸽精准推送服务

- 标签推送
- A/B 推送
- 用户分群推送
- 智能精准推送

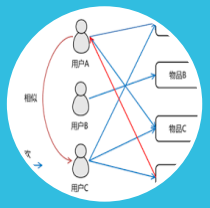
Hybrid Model



$$P = \frac{1}{1 + e^{-u}}$$

$$u = b_0 + \sum_{i=1}^{n}(x_i w_i)$$

规则引擎

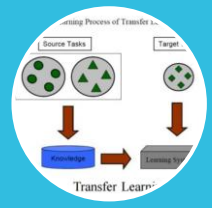


协同过滤

预估模型

深度学习

在线学习



迁移学习

精准推送服务基本流程

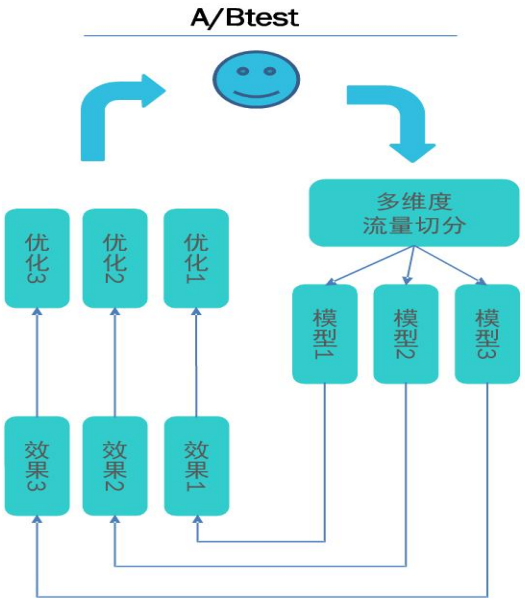
算法选择
- 目标
- 场景

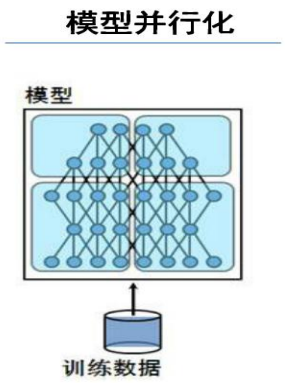
模型训练
- 样本
- 特征
- FTRL

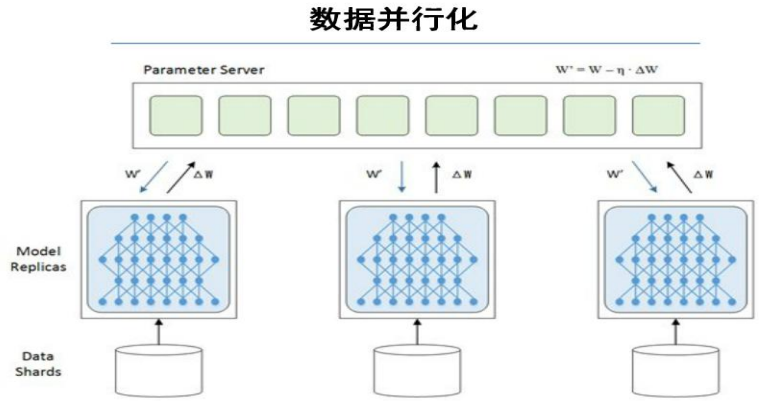
模型评估
- AUC
- 准确率
- 召回率

预测
- 单模型
- 混合模型
- 迁移学习

效果统计
- CTR
- A/B Test







丰富的产品线和数据积累

月活跃用户超 8 亿
在线人际关系链超 1000 亿

注册用户数超 10 亿
月活跃超 7 亿

活跃用户数超 6 亿
智能终端活跃用户数超 5 亿

移动支付超过 5 亿笔

其他: 新闻, 视频, 音乐, 游戏, 手机管家

账号信息
年龄, 性别, 学历...

UGC
QZone日志, 图像, 评论...

社交属性
QQ, Qzone, 好友关系链...

金融
理财, 消费, 信用...

行为偏好
新闻, 视频, 公众号...

商业兴趣
广告, 商品, 品牌...

游戏爱好
游戏时长, 类型, 付费...

连接

账号

PGM

分析

DNN

LPA

标准化

结构化

非结构

分布式计算、机器学习基础支撑平台

业务在线系统

界面化、可视化

任务调度/工作流　　黄金眼　　IDE

算法应用平台

Tesla 图像处理

TDBank

DATA API

多维实时检索 Hermes

CKV 系统 TDE

离线计算 MR

实时计算 TRC

机器学习 Angel (参数服务) Mariana

精准推荐 FTRL LR

AI & BI OCR NLP

DMP LookAlike

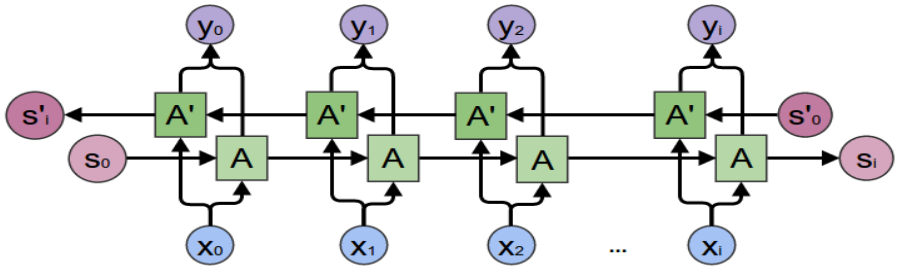
资源、任务调度 GaiaStack、Lhost

CPU 集群

GPU 集群

Storage HDFS CEPH

机器学习可视化界面操作



for $t = 1$ to T **do**
  Forward pass for the forward hidden layer, storing activations at each timestep
for $t = T$ to 1 **do**
  Forward pass for the backward hidden layer, storing activations at each timestep
for all t, in any order **do**
  Forward pass for the output layer, using the stored activations from both hidden layers

for all t, in any order **do**
  Backward pass for the output layer, storing δ terms at each timestep
for $t = T$ to 1 **do**
  BPTT backward pass for the forward hidden layer, using the stored δ terms from the output layer
for $t = 1$ to T **do**
  BPTT backward pass for the backward hidden layer, using the stored δ terms from the output layer

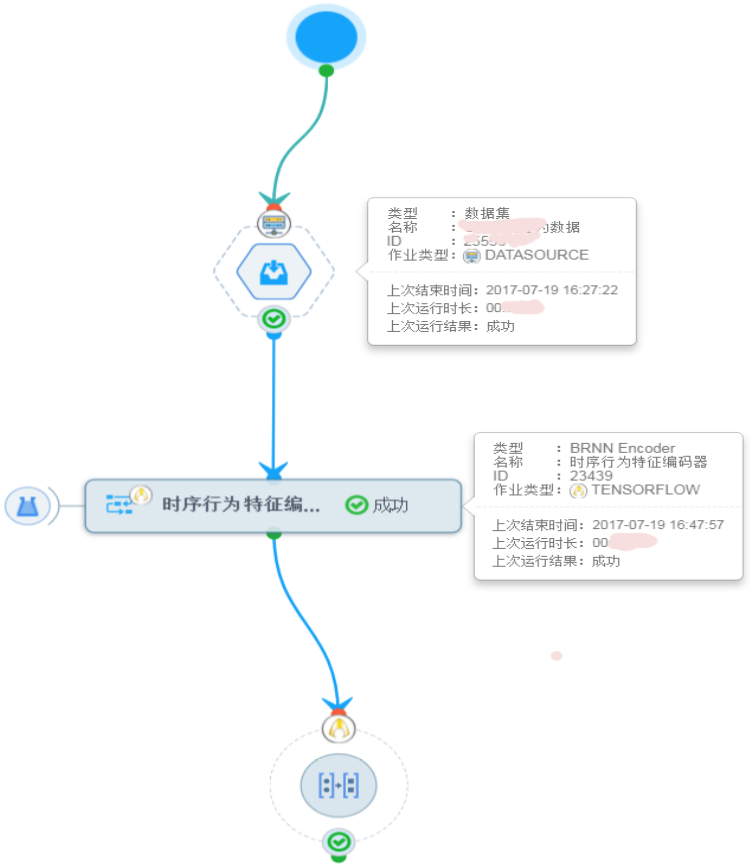
$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1}$$

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}$$
$$b_h^t = \theta(a_h^t)$$

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}$$

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t$$

点击行为预测：BRNN（TESLA ML）



信鸽完整技术解决方案

平台共赢

泛推送
- 公众号
- 短信

Gateway
- 数据开放
- 平台对接

运营平台
- 立体化监控
- 数据透视
- 画像分析
- 配置管理

场景营销

- 地理围栏推送
- 推送效果再营销
- 行为场景推送
- 行业热词模版

知识发现
用户聚类
社群挖掘

基础平台工具
- Angel
- 云控系统
- Gaia
- TDW/Storm

精准

- 标签定向推送
- A/B Test 推送
- MTA 用户分群、事件数据互通

模型算法
- 流失预测
- FTRL

公共组件
- 全局ID分配
- 位图标签系统
- 多维实时检索
- 限频安全旁路
- CKV

时效

- 分布式检索推送
- 智能选路
- 分级推送调度
- 实时效果跟踪
- 通道隔离
- 预推送

抵达

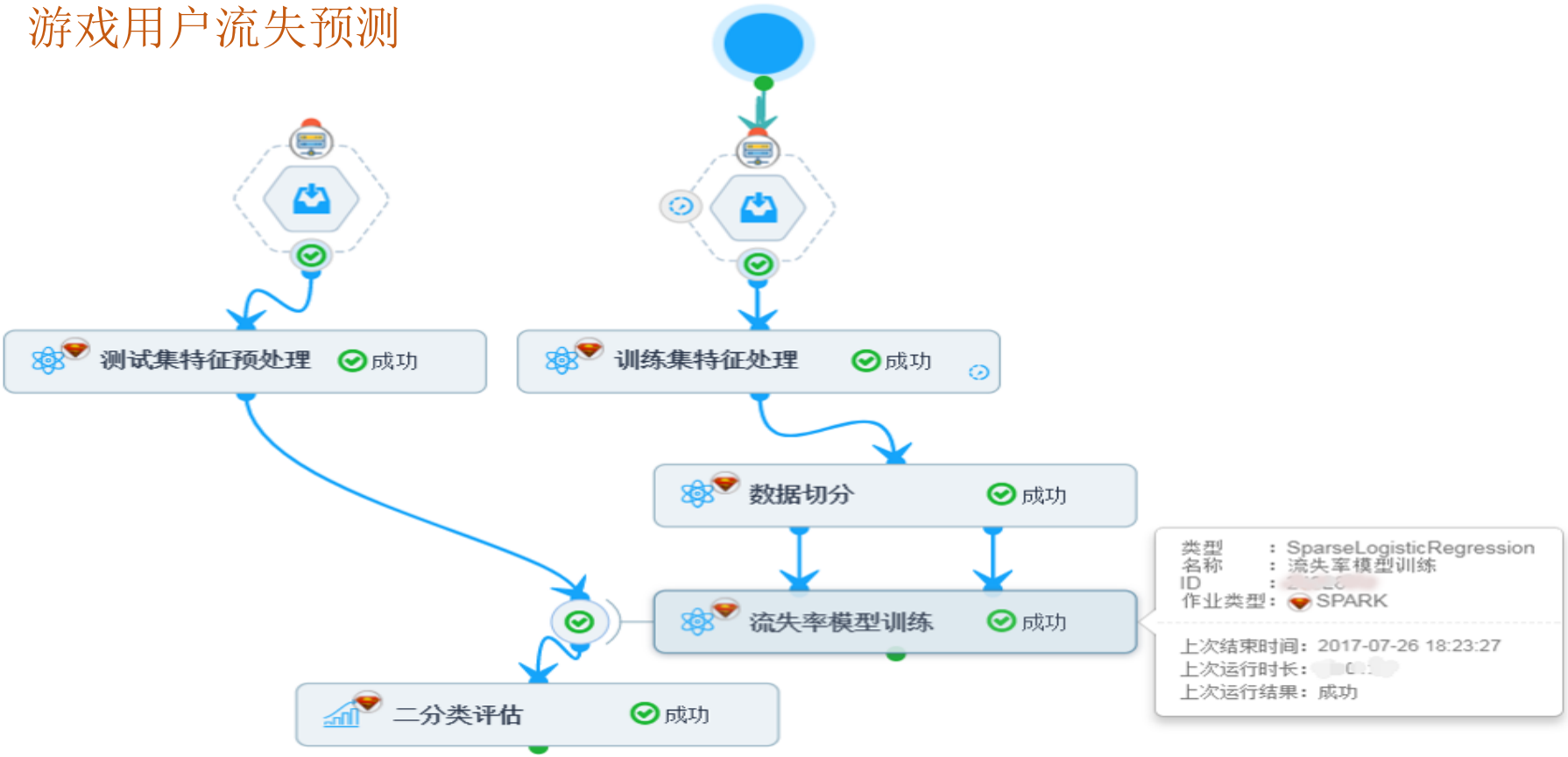
厂商通道
- 小米通道
- 华为通道
- OPPO通道

APNS
FCM

自建通道
- Android iOS
- 智能硬件 物联网

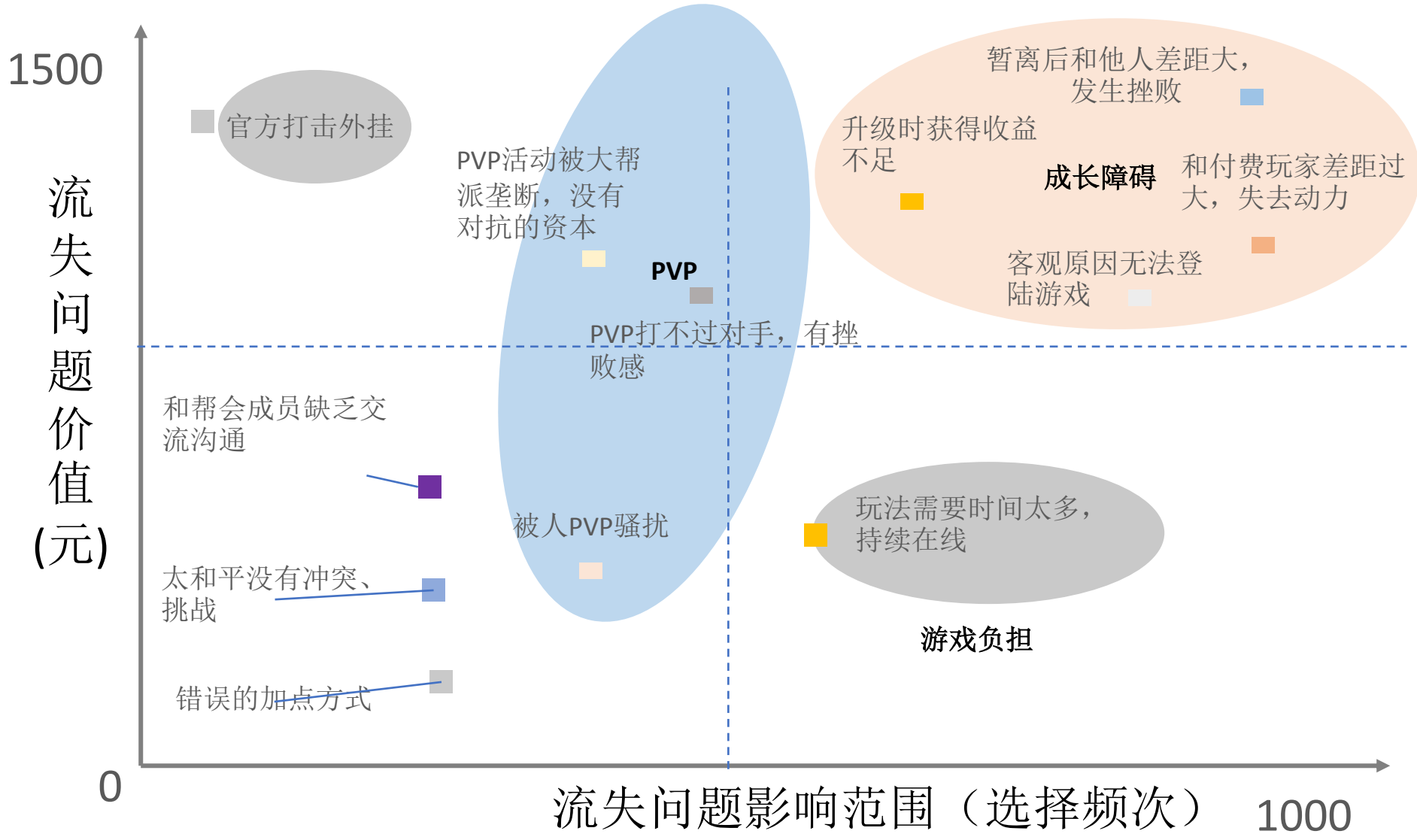
一个案例：提升游戏留存

游戏用户流失预测



一个案例：提升游戏留存

用户流失归因分析



一个案例：提升游戏应用留存

信鸽

移动推送服务

用户分群

缺少挑战

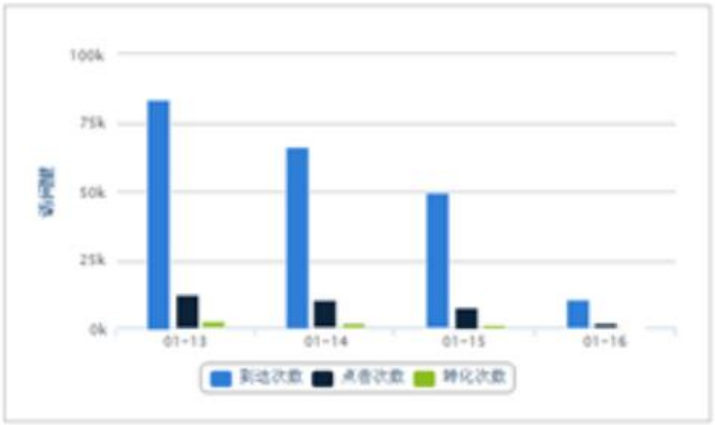
PVP失利

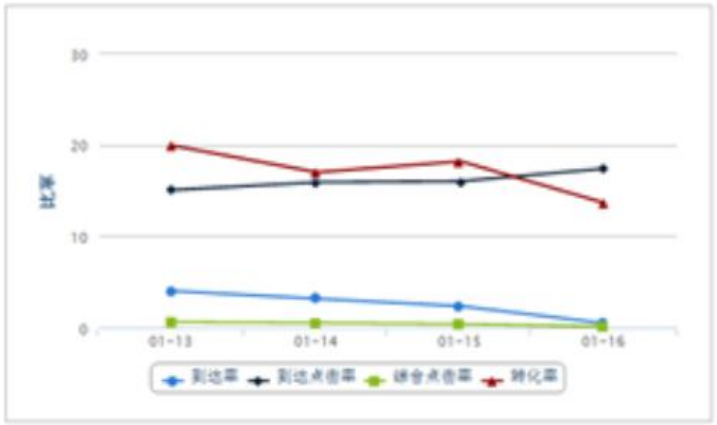
用户属性

消息精准推送

我有残局，你敢来挑战吗？

生日独享大礼包





KP（开放生态、平台合作）

- 微信开放平台
- 腾讯云
- 腾讯开放平台
- 应用宝
- APICloud
- 易起秀

Customer Segments

- 移动开发者
- 媒体
- 移动运营者
- 大数据从业者
- Nielsen

效率

成本

增值

易用

KR

- 用户、设备画像
- 协同营销
- 大数据解决方案
- 流量
- 资本

Key Activities

- Crash 监控
- 可视化埋点
- 用户分群
- 渠道跟踪
- 精准推送
- 反作弊

Value Proposition

- SaaS
- AI
- BI

TBDS　DIX　OCR

参考资料：

1. 《Systems Performance Enterprise and the Cloud.2013》

Thanks