Tencent Big Data

腾讯信鸽实时推送演进与实践

甘恒通
腾讯大数据高级工程师

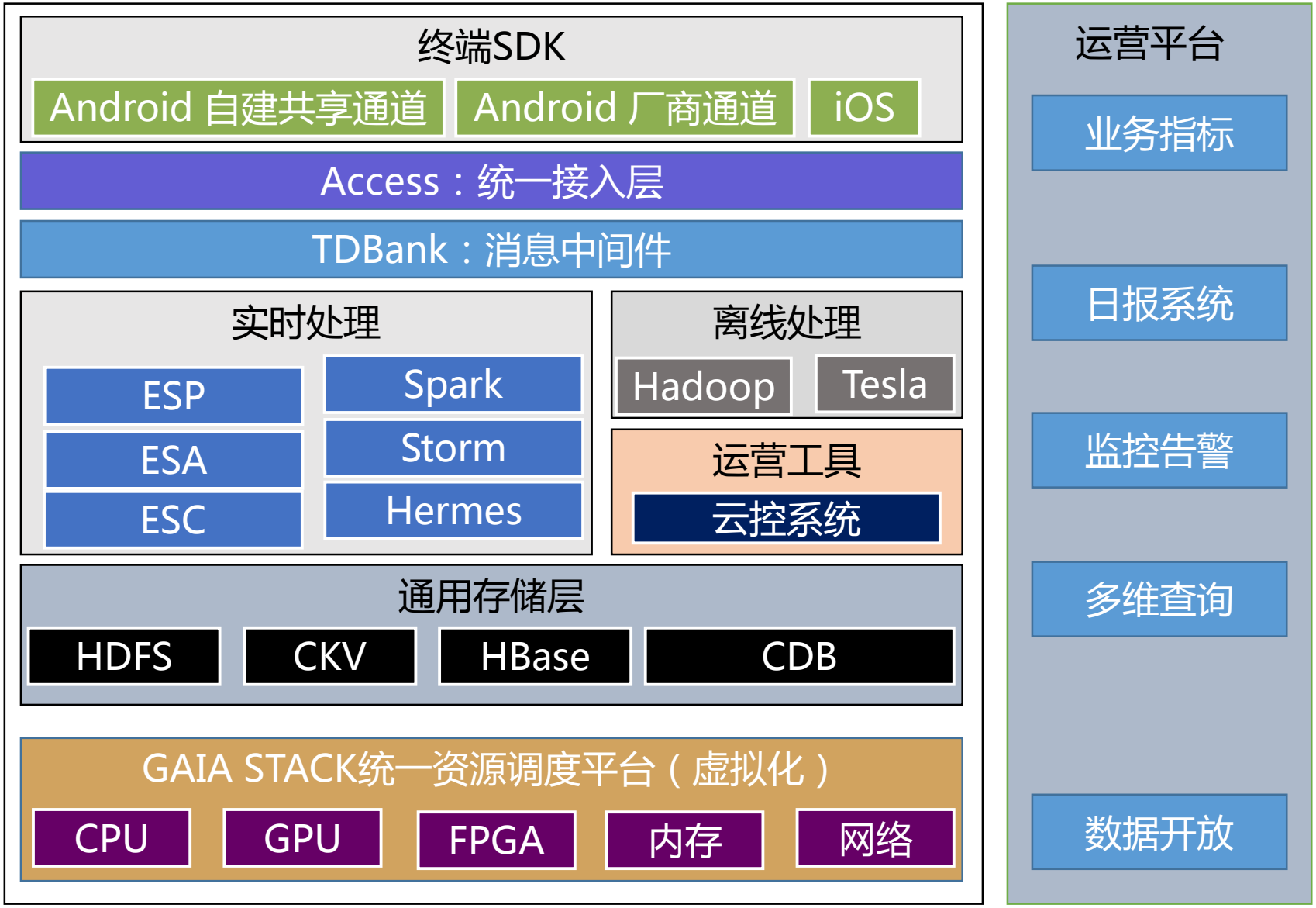
数据平台部

目录

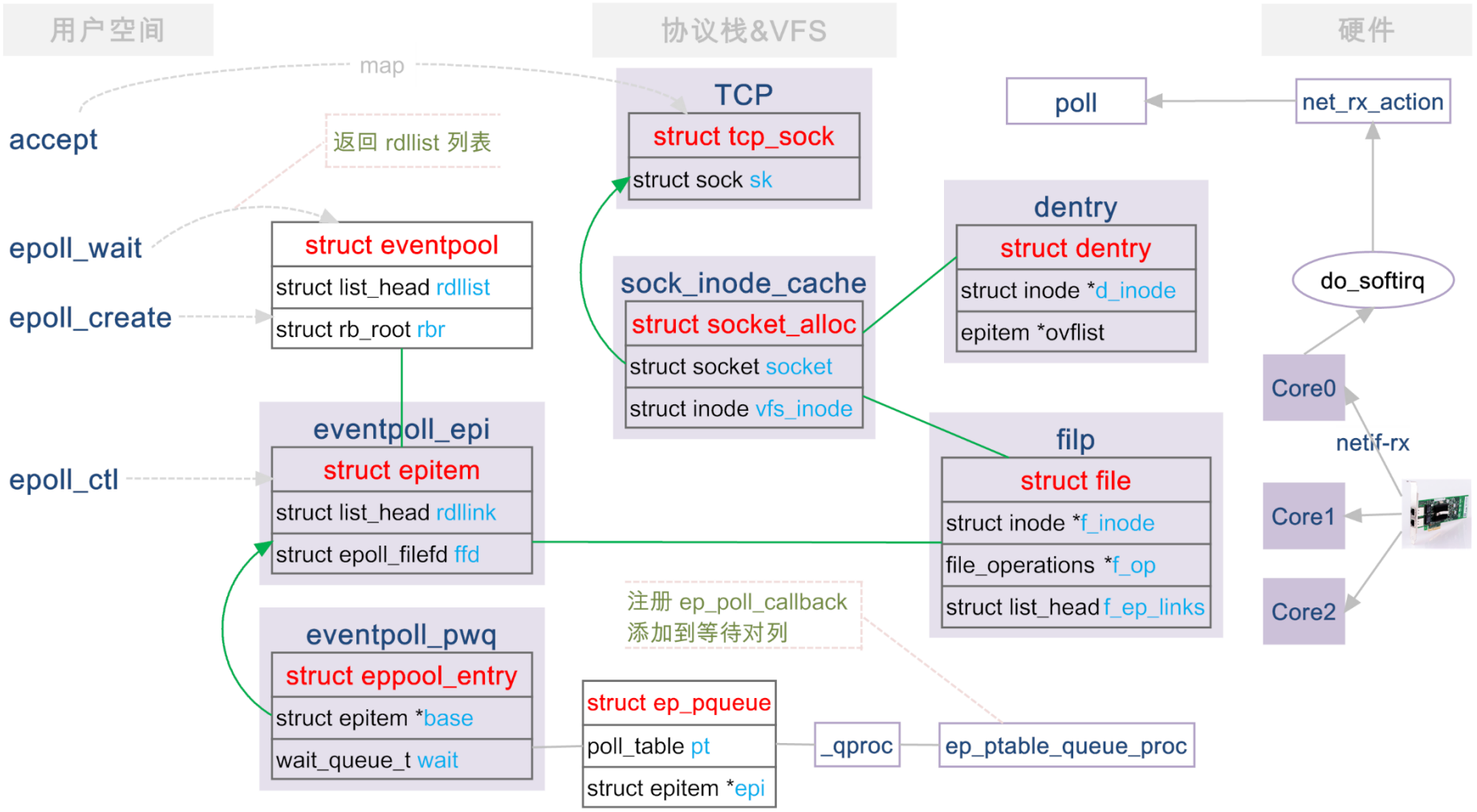
信鸽推送的挑战

- 海量
  - 终端长连接
  - 推送量
- 实时
  - 1KW/s
- 精准
  - 画像/机器学习人群分类

信鸽推送系统解决方案


终端SDK
Android 自建共享通道
Android 厂商通道
iOS
Access：统一接入层
TDBank：消息中间件
实时处理
ESP
ESA
ESC
Spark
Storm
Hermes
离线处理
Hadoop
Tesla
运营工具
云控系统
通用存储层
HDFS
CKV
HBase
CDB
GAIA STACK统一资源调度平台（虚拟化）
CPU
GPU
FPGA
内存
网络
运营平台
业务指标
日报系统
监控告警
多维查询
数据开放

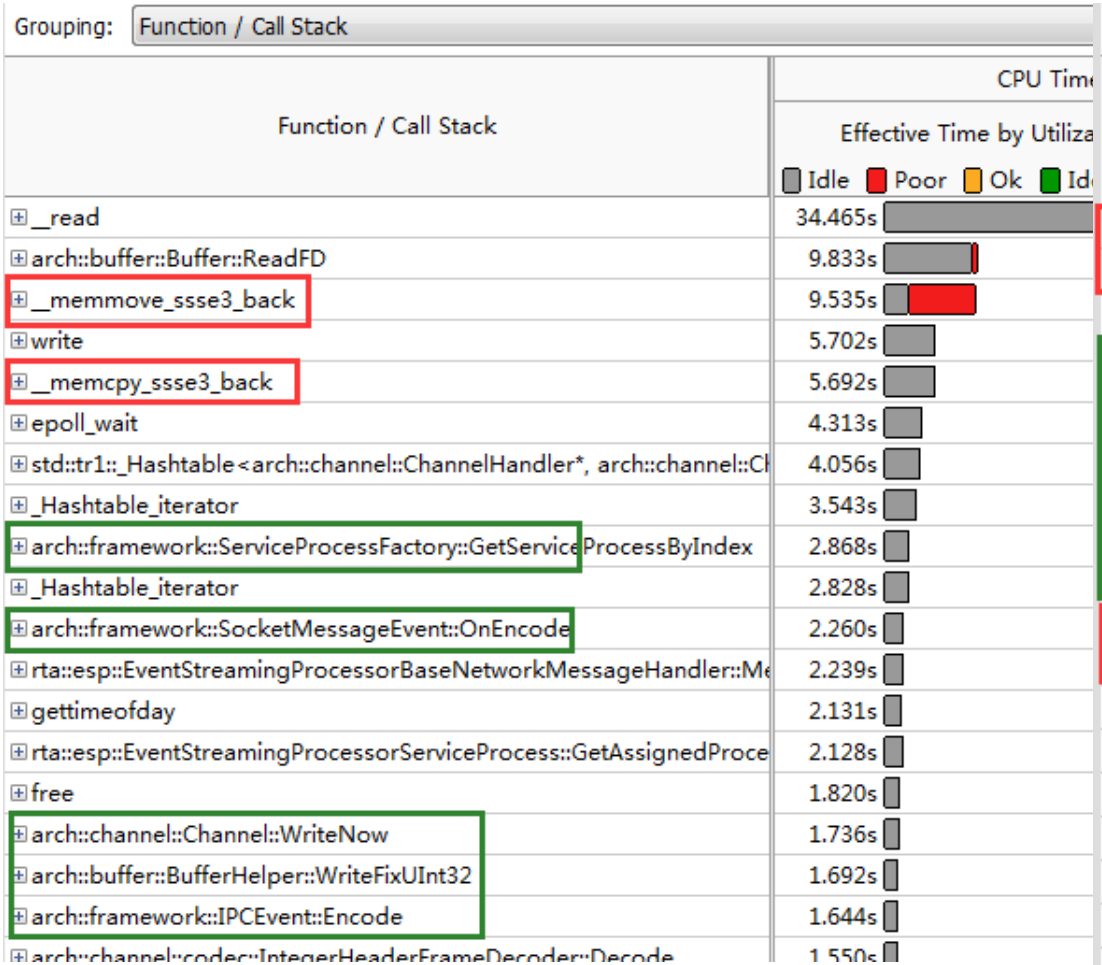
单机性能优化--关键环节

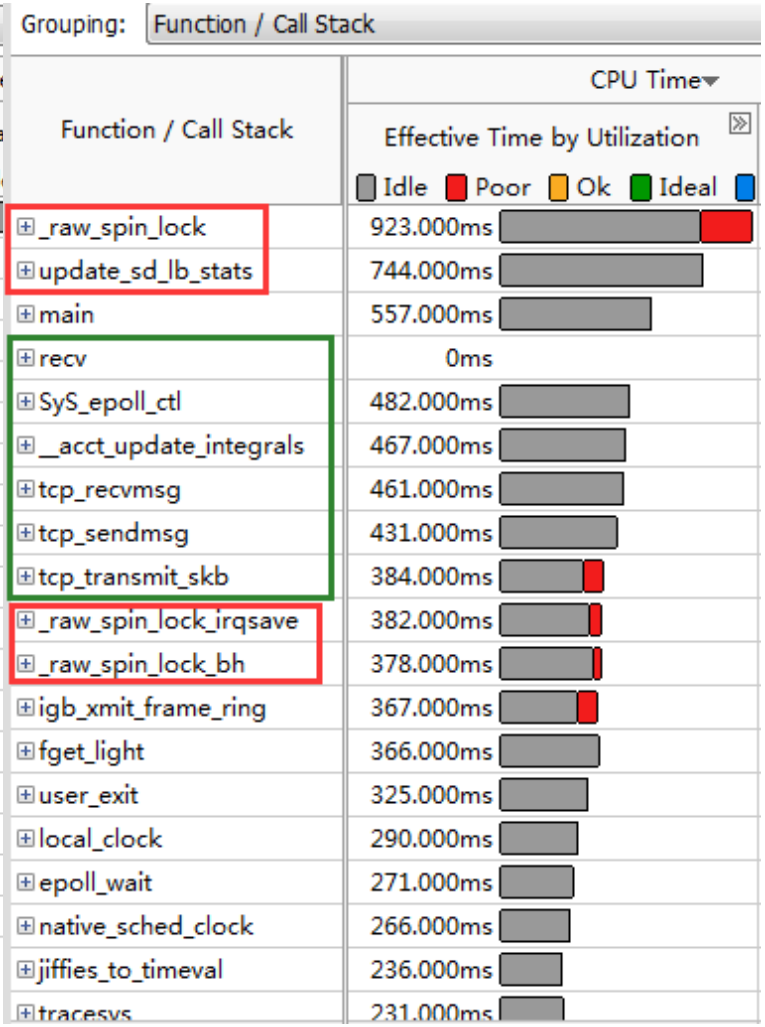

用户空间
协议栈&VFS
硬件
map
accept
返回 rdllist 列表
epoll_wait
epoll_create
epoll_ctl
struct eventpool
struct list_head rdllist
struct rb_root rbr
eventpoll_epi
struct epitem
struct list_head rdllink
struct epoll_filefd ffd
eventpoll_pwq
struct eppool_entry
struct epitem *base
wait_queue_t wait
TCP
struct tcp_sock
struct sock sk
sock_inode_cache
struct socket_alloc
struct socket socket
struct inode vfs_inode
dentry
struct dentry
struct inode *d_inode
epitem *ovflist
filp
struct file
struct inode *f_inode
file_operations *f_op
struct list_head f_ep_links
注册 ep_poll_callback
添加到等待对列
struct ep_pqueue
poll_table pt
struct epitem *epi
_qproc
ep_ptable_queue_proc
poll
net_rx_action
do_softirq
Core0
netif-rx
Core1
Core2

单机性能优化--操作系统--配置部分

配置类型	配置项	值
系统打开文件资源限制	/proc/sys/fs/file-max	2500000
进程打开文件句柄限制	/proc/sys/fs/nr_open	2500000
Epoll监听fd个数限制	/proc/sys/fs/epoll/max_user_watches	2500000
内核分配给TCP内存	/proc/sys/net/ipv4/tcp_mem	754752 1006336 1509504
Listen fd backlog	/proc/sys/net/core/somaxconn	1024
fast recycling of TWsocketsv	/proc/sys/net/ipv4/tcp_tw_recycle	1
Reuse TW socket	/proc/sys/net/ipv4/tcp_tw_reuse	1
拥塞控制算法	/proc/sys/net/ipv4/tcp_westwood	0
Eth scatter-gather	Eth sg	Off
Eth generic-segmentation-offload	Eth gso	Off

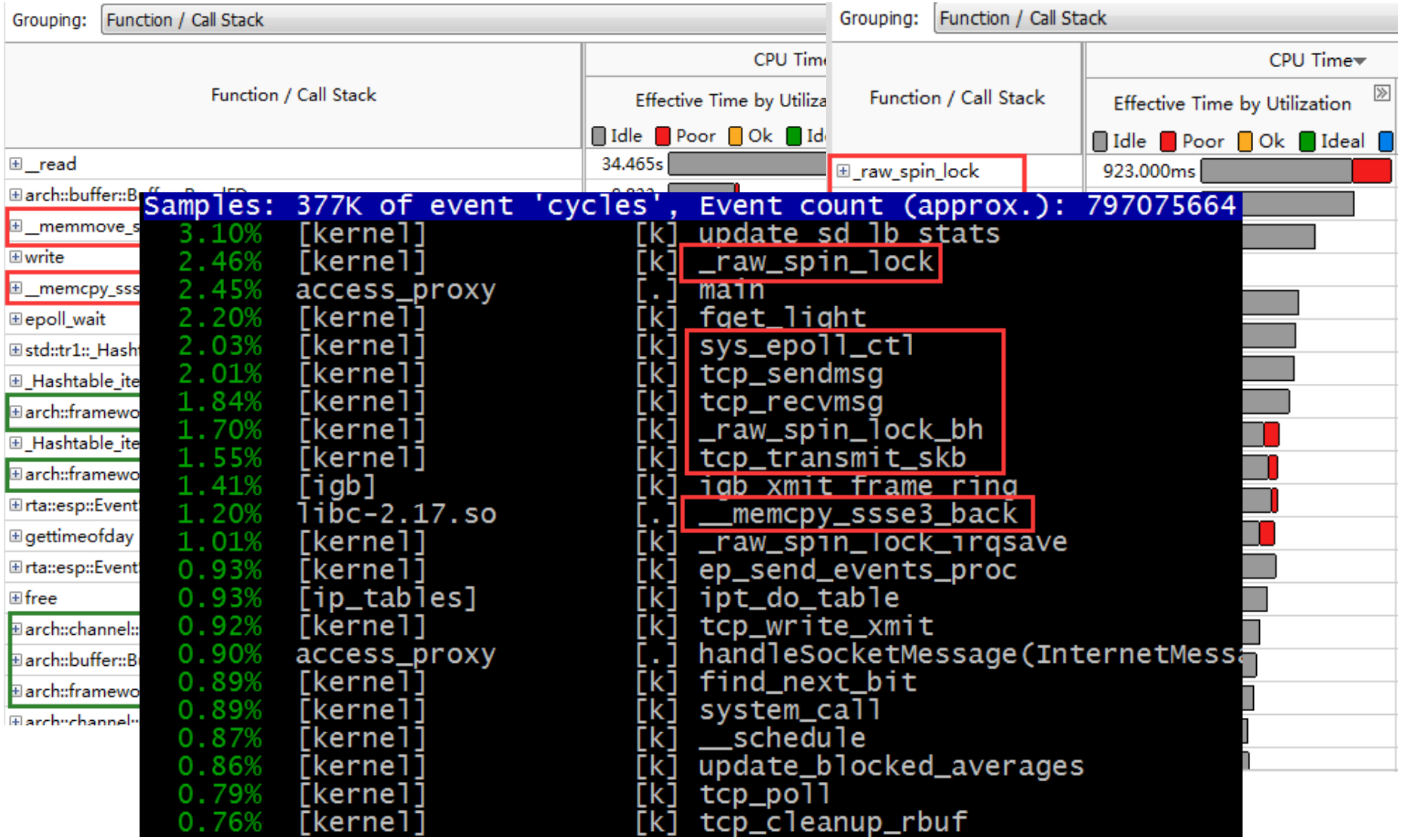
单机性能优化--性能评测





Intel® VTune™ Amplifier

单机性能优化--性能评测


Grouping: Function / Call Stack
Function / Call Stack
CPU Time
Effective Time by Utilization
Idle Poor Ok Ideal
_read 34.465s
_raw_spin_lock 923.000ms
write
epoll_wait
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(InternetMess
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

单机性能优化--硬件性能挖掘

```
int check_support_sse4_2() {
    int res=0;
    __asm__ __volatile__(
                        "movl $1,%%eax\n\t"
                        "cpuid\n\t"
                        "test $0x0100000,%%ecx\n\t"
                        "jz 1f\n\t"
                        "movl $1,%0\n\t"
                        "1:\n\t"
                        :"=m"(res)
                        :
                        :"eax","ebx","ecx","edx");
    return res;
}
```

SSE4.2检测

unsigned int _mm_crc32_u8(...)
unsigned int _mm_crc32_u16(...)
unsigned int _mm_crc32_u32(...)
unsinged __int64 _mm_crc32_u64(...)

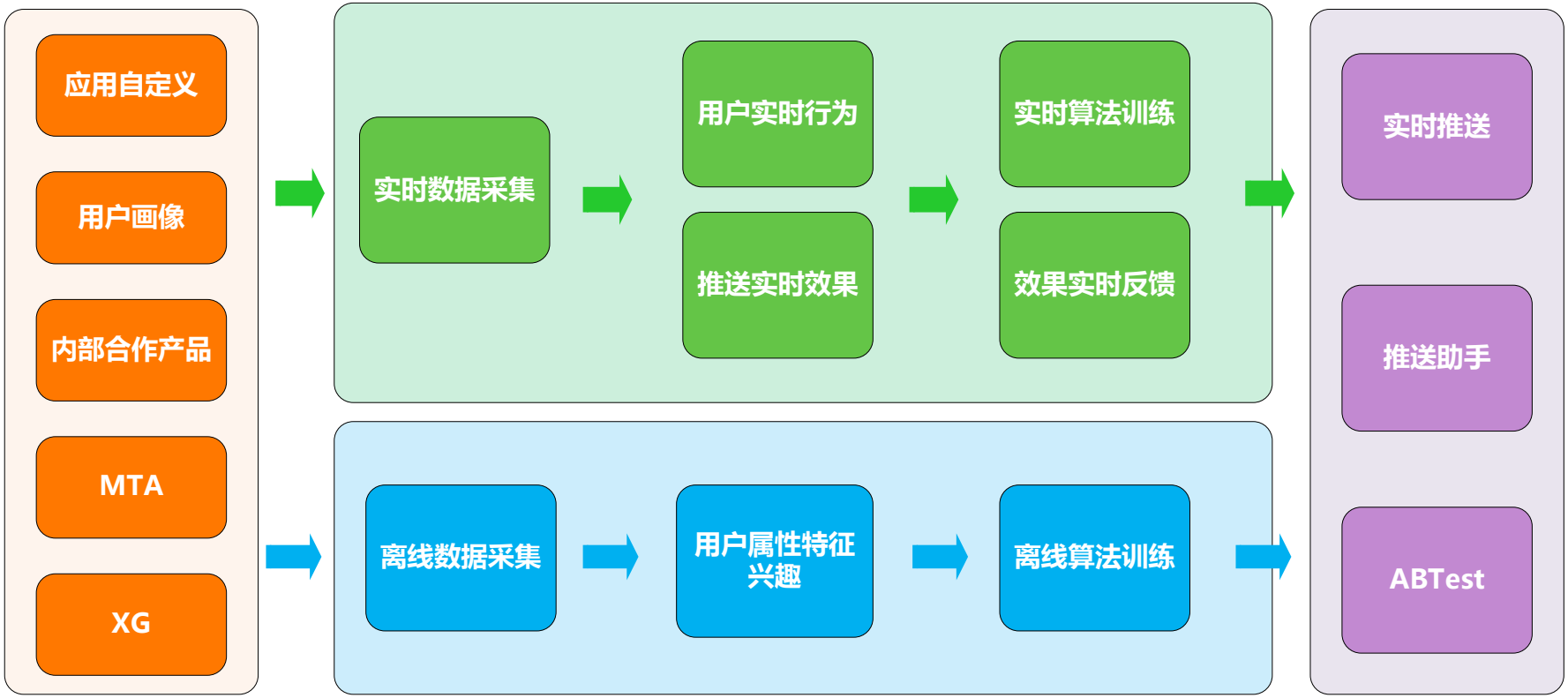
SSE4.2 CRC32支持

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

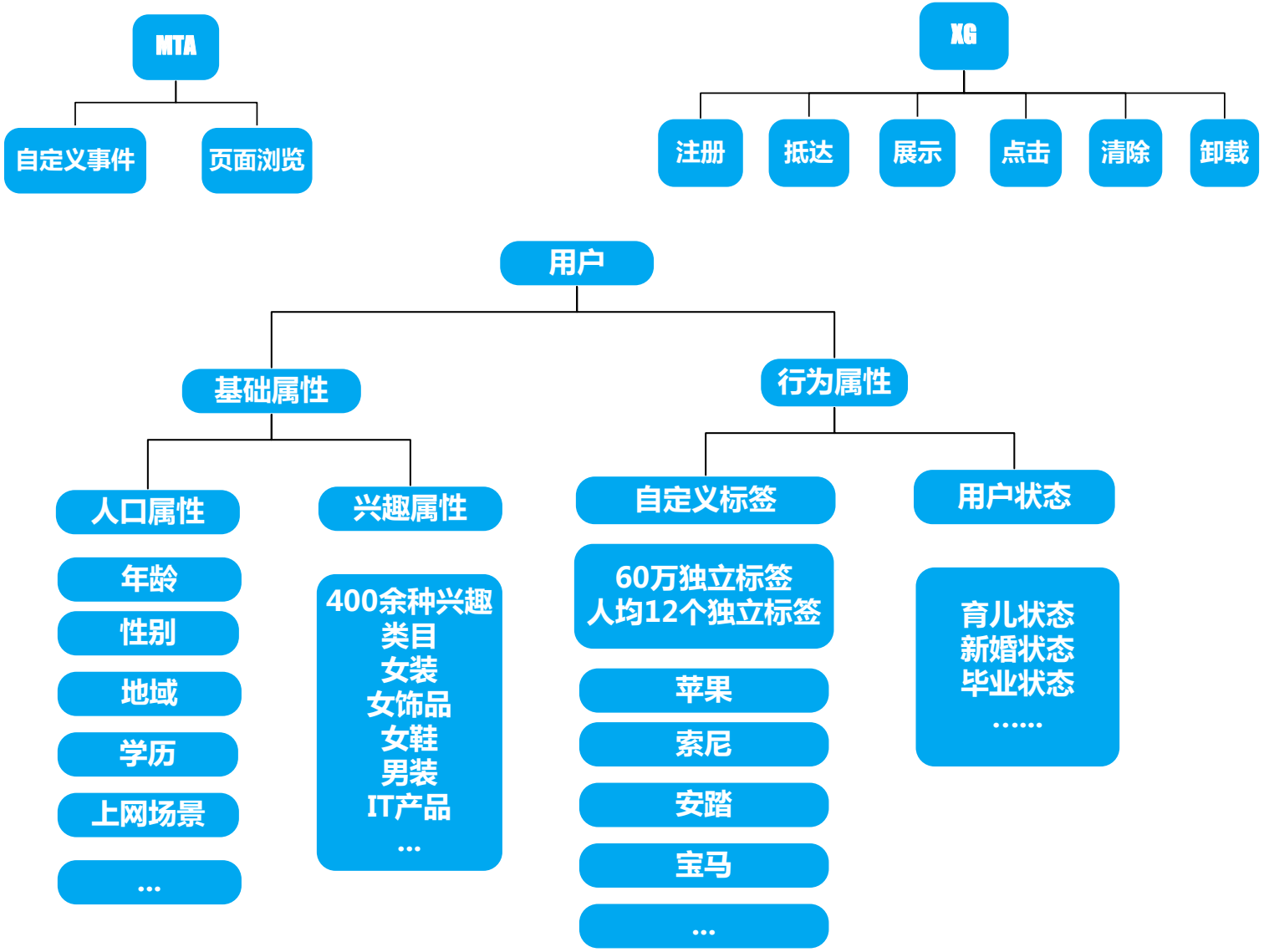
__errno_location	0ms	5,060,000
update_cfs_shares	20.000ms	15,180,000
retransmits_timed_out	20.000ms	32,890,000
MD::BinUtil::crc32csse	20.000ms	43,010,000
ipv4_dst_check	20.000ms	43,010,000
ip_output	20.000ms	50,600,000
inode_init_once	20.000ms	22,770,000

Intel SSE4.2 CRC32与常规CRC32性能对比
执行 35W 次计算结果（SSE 快接近 **6** 倍）

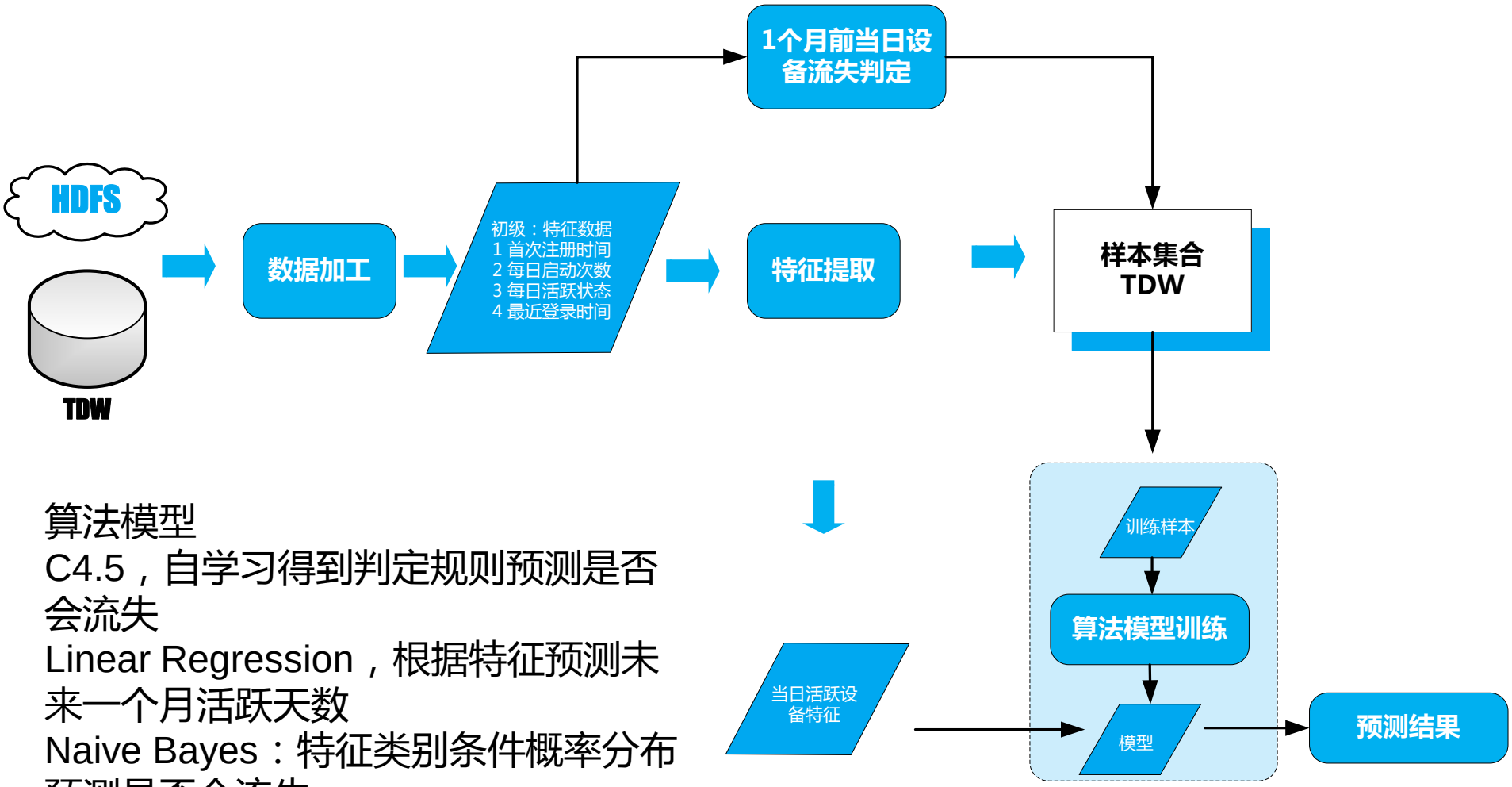
实时精准推送系统


应用自定义
用户画像
内部合作产品
MTA
XG
实时数据采集
用户实时行为
推送实时效果
实时算法训练
效果实时反馈
离线数据采集
用户属性特征
兴趣
离线算法训练
实时推送
推送助手
ABTest

数据


MTA
自定义事件
页面浏览
XG
注册
抵达
展示
点击
清除
卸载
用户
基础属性
行为属性
人口属性
兴趣属性
自定义标签
用户状态
年龄
性别
地域
学历
上网场景
...
400余种兴趣
类目
女装
女饰品
女鞋
男装
IT产品
...
60万独立标签
人均12个独立标签
苹果
索尼
安踏
宝马
...
育儿状态
新婚状态
毕业状态
.....

人群挖掘



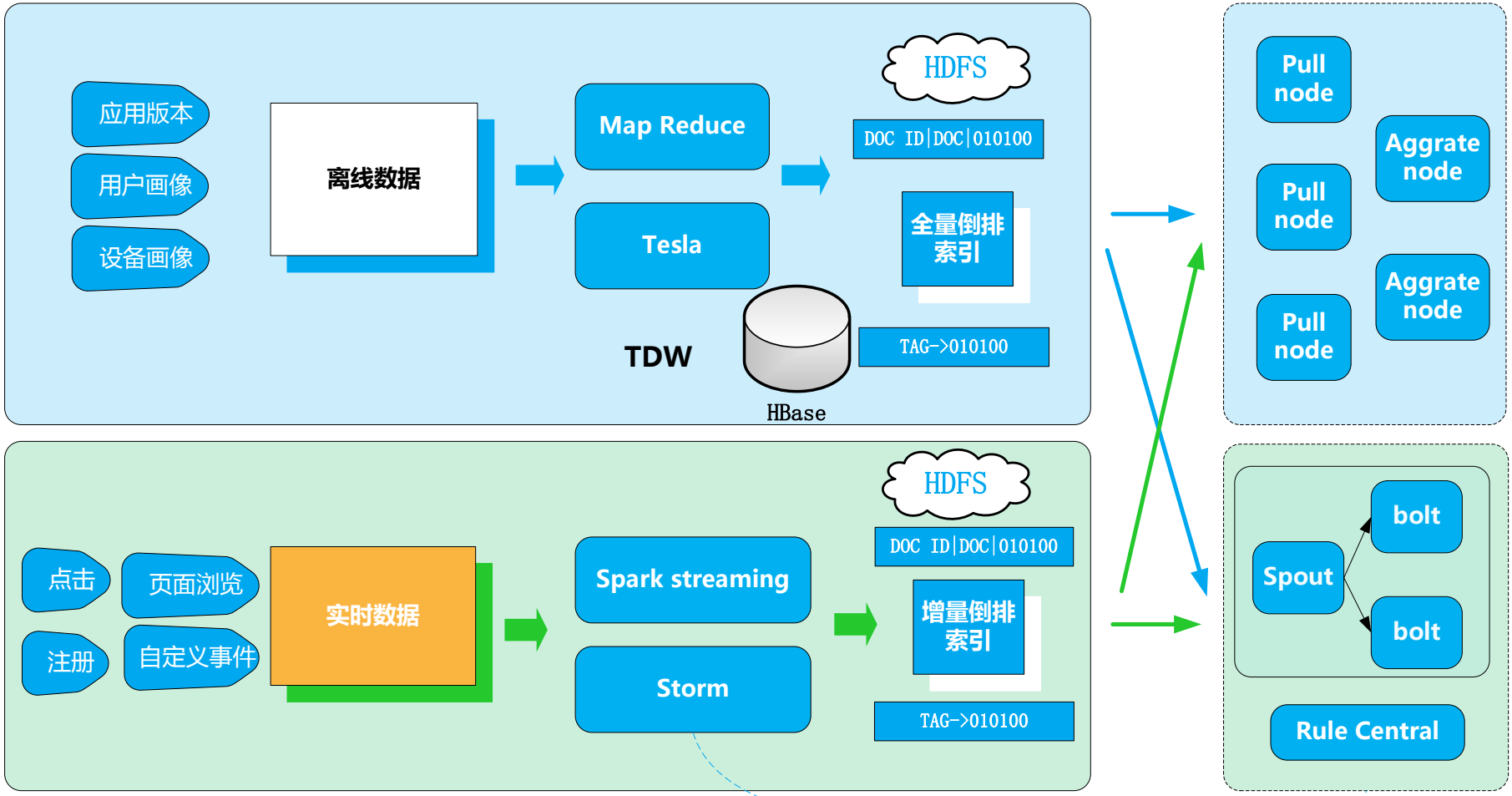
算法模型

C4.5，自学习得到判定规则预测是否会流失

Linear Regression，根据特征预测未来一个月活跃天数

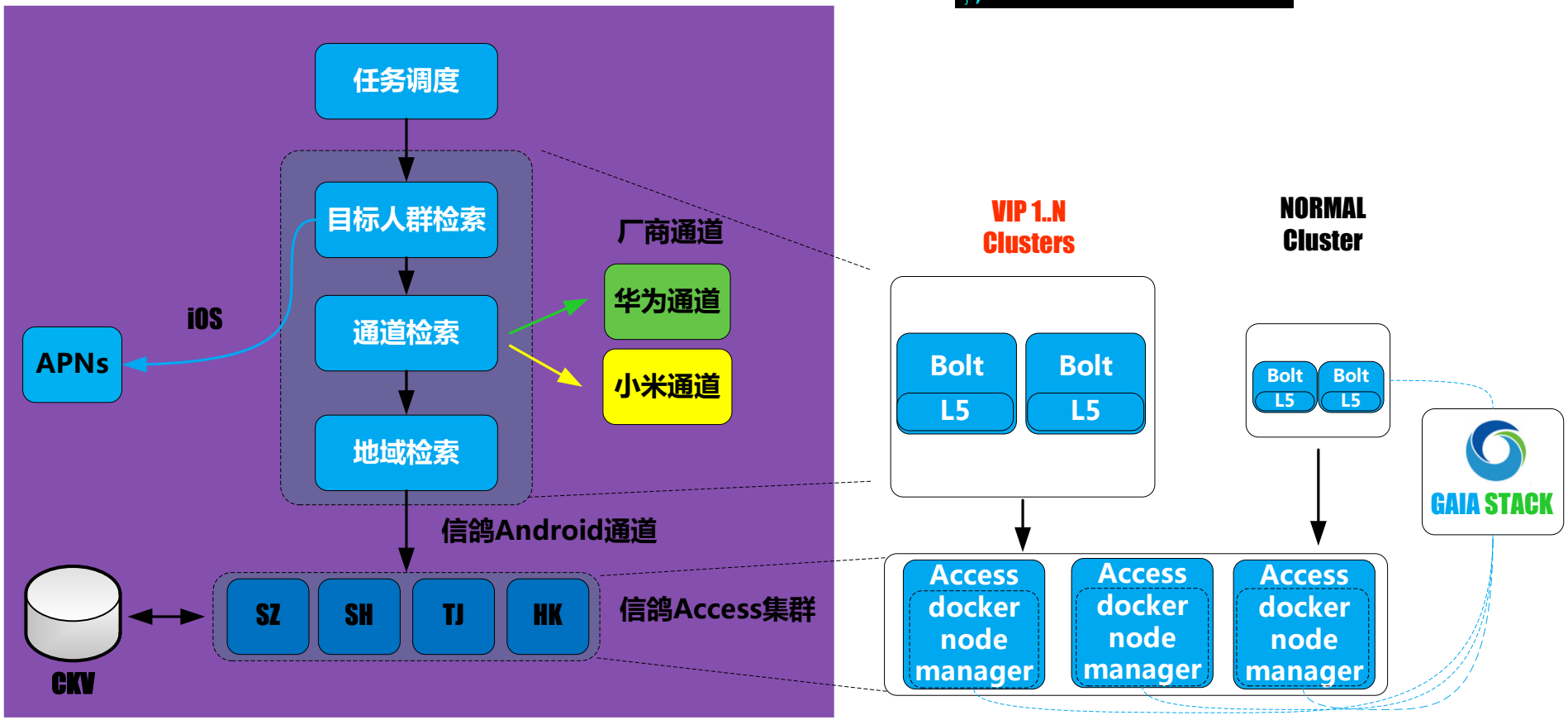
Naive Bayes：特征类别条件概率分布预测是否会流失

实时推送&多维分析


应用版本
用户画像
设备画像
离线数据
Map Reduce
Tesla
TDW
HDFS
DOC ID|DOC|010100
全量倒排索引
TAG->010100
HBase
Pull node
Pull node
Pull node
Aggrate node
Aggrate node
点击
页面浏览
注册
自定义事件
实时数据
Spark streaming
Storm
HDFS
DOC ID|DOC|010100
增量倒排索引
TAG->010100
Spout
bolt
bolt
Rule Central

实时推送

```
union {
    uint64_t    ullFlow;
    struct router {
        uint8_t ucRegion;
        uint8_t ucSetID;
        uint8_t ucBoxID;
        uint8_t ucNodeID;
        uint32_t uiSockIdx;
    };
};
```



效果评估

目标页面集合

自定义参数事件

</code>

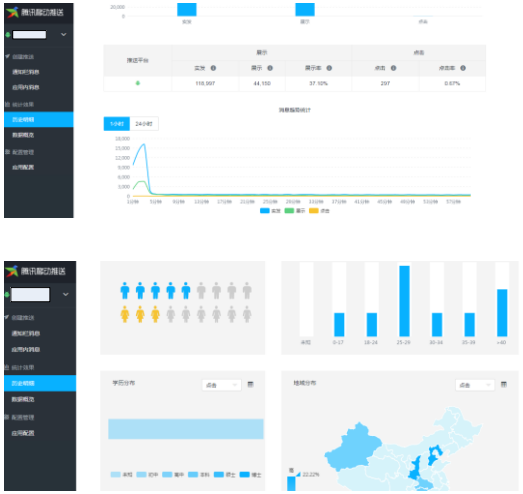
数据源

点击 | 自定义事件 | 注册 | 用户画像 | 浏览 | 设备画像

Storm

Redis

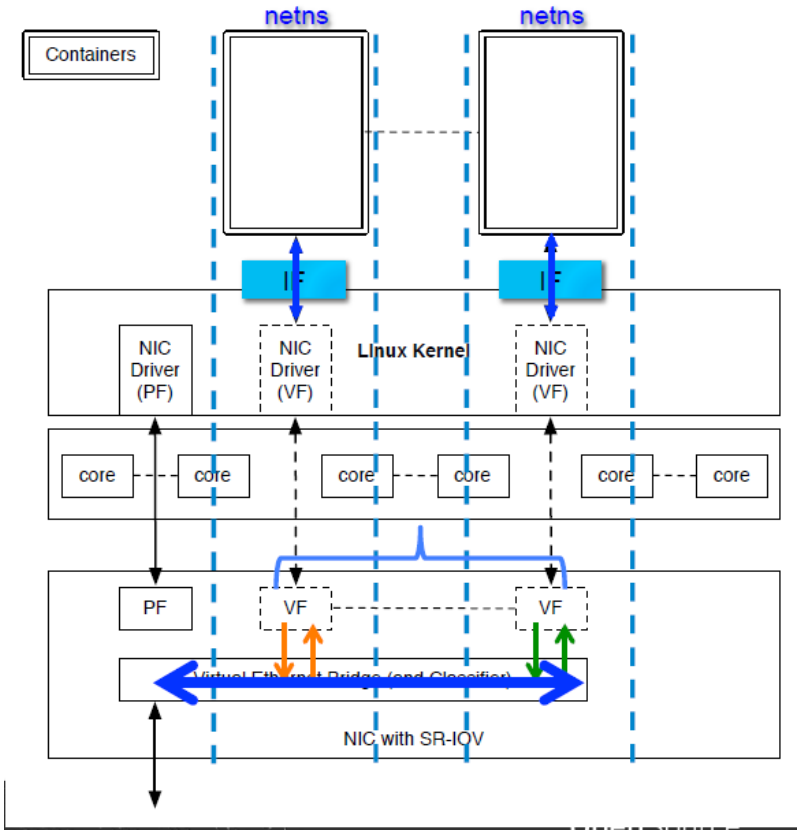
Proxy



Extract → Filter → Transform → Load → Update tag

Parser → Compute

软硬虚拟化



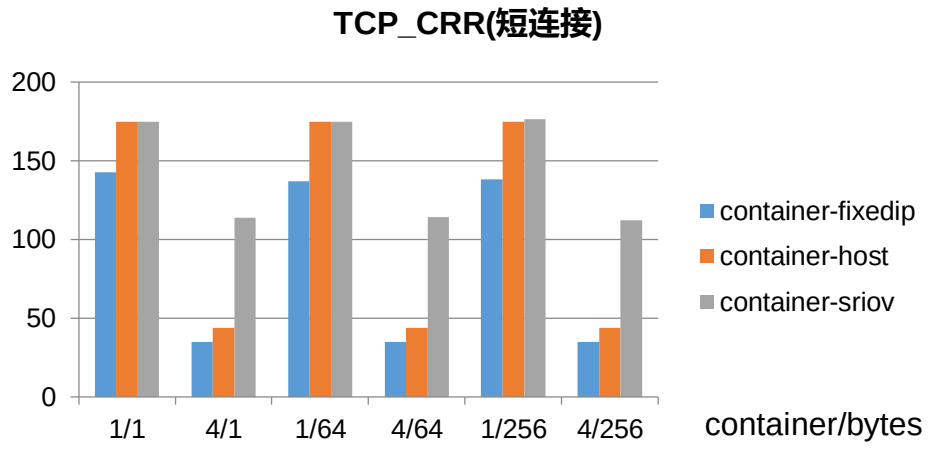
isolcpus:

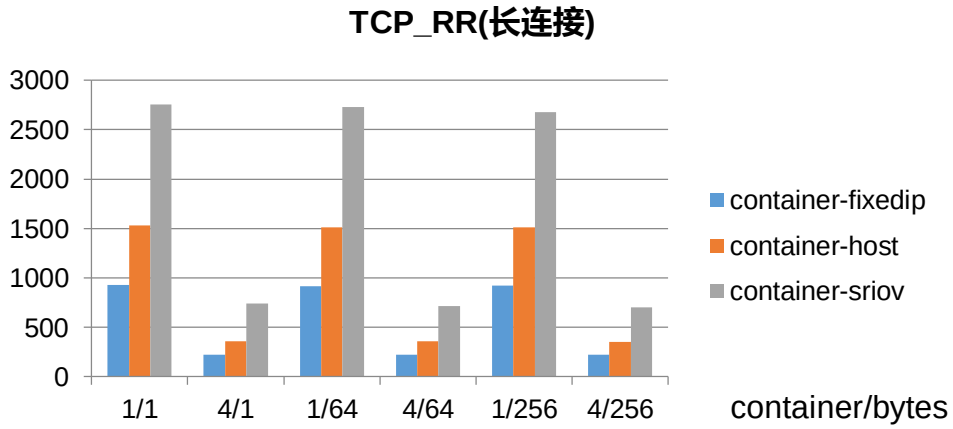
- Boot-time (Kernel boot parameter)
  - `... default_hugepagesz=1G ... isolcpus=12-15`
- Isolation from timers from other CPUs.

Cgroups/cpuset.cpus:

- Run-time
  - Isolate target CPUs (Next Page)
  - Run Container on those CPUs
- Same as isolcpus except the hrtimer issue

```
$ docker run -ti --cpuset-cpus="12-15" ...
```





腾讯云分析

腾讯大数据

信鸽推送

技术支持	商务咨询
dtsupport@tencent.com	data@tencent.com