阿里网络故障智能化治理的实践:故障自动发现&恢复

何源（荆杭）
阿里巴巴高级技术专家

体量大

型号多&架构多

结构复杂

网络自身依赖


druid


Flink


APACHE
kafka


elasticsearch


APACHE
HBASE


mongoDB


APACHE
Spark


hadoop

采集

故障发现

根因定位

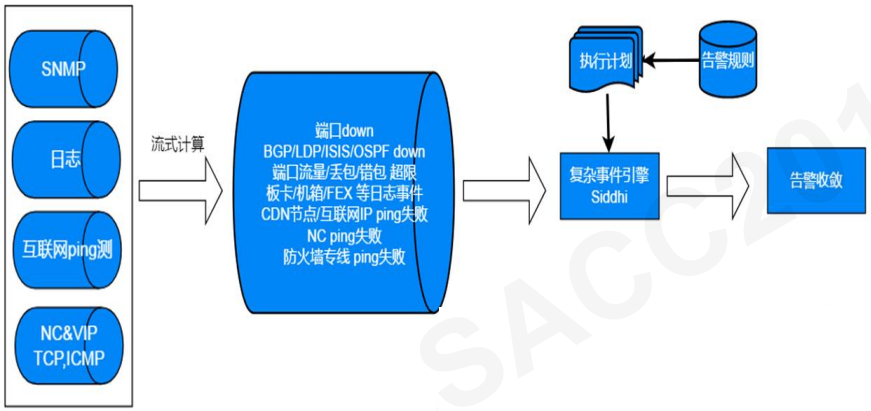
自动化故障处理

网络设备/防火墙/ANAT/LVS

互联网质量

专线&VIP

服务器丢包&延时



设备状态扫描

变更扫描

流量攻击扫描

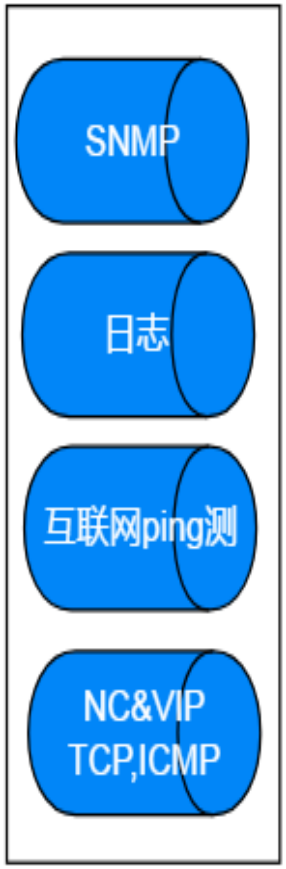
运营商状态扫描

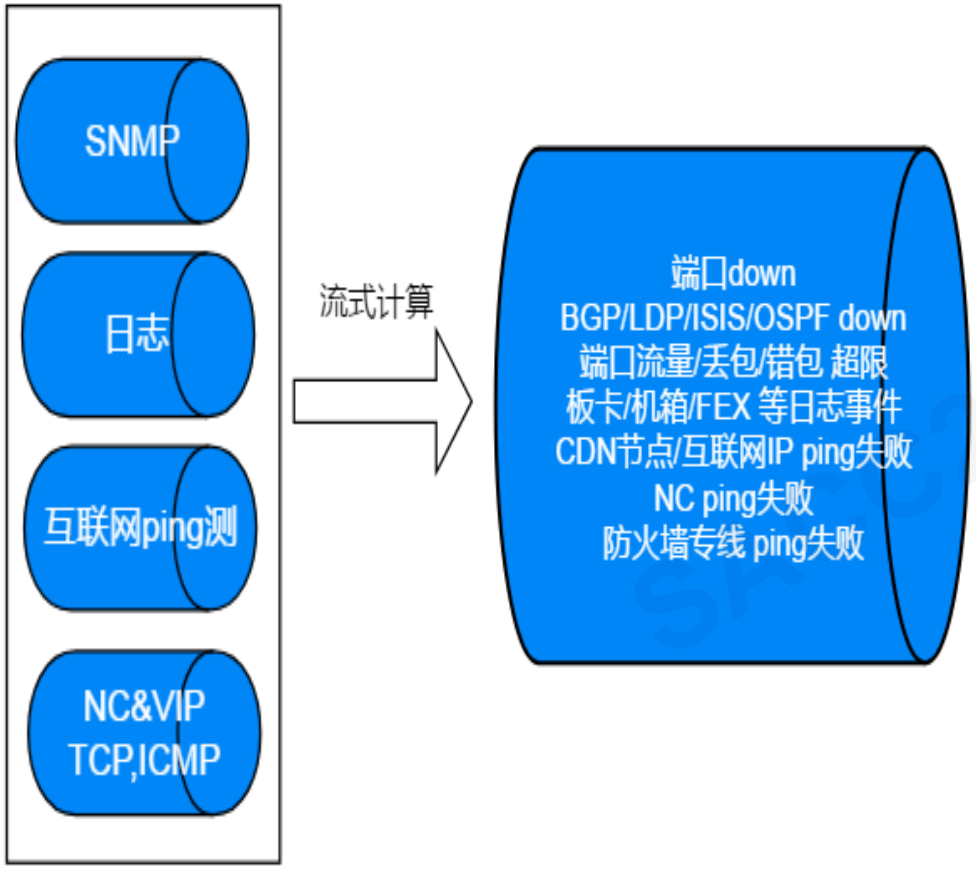
端口/板卡/整机隔离

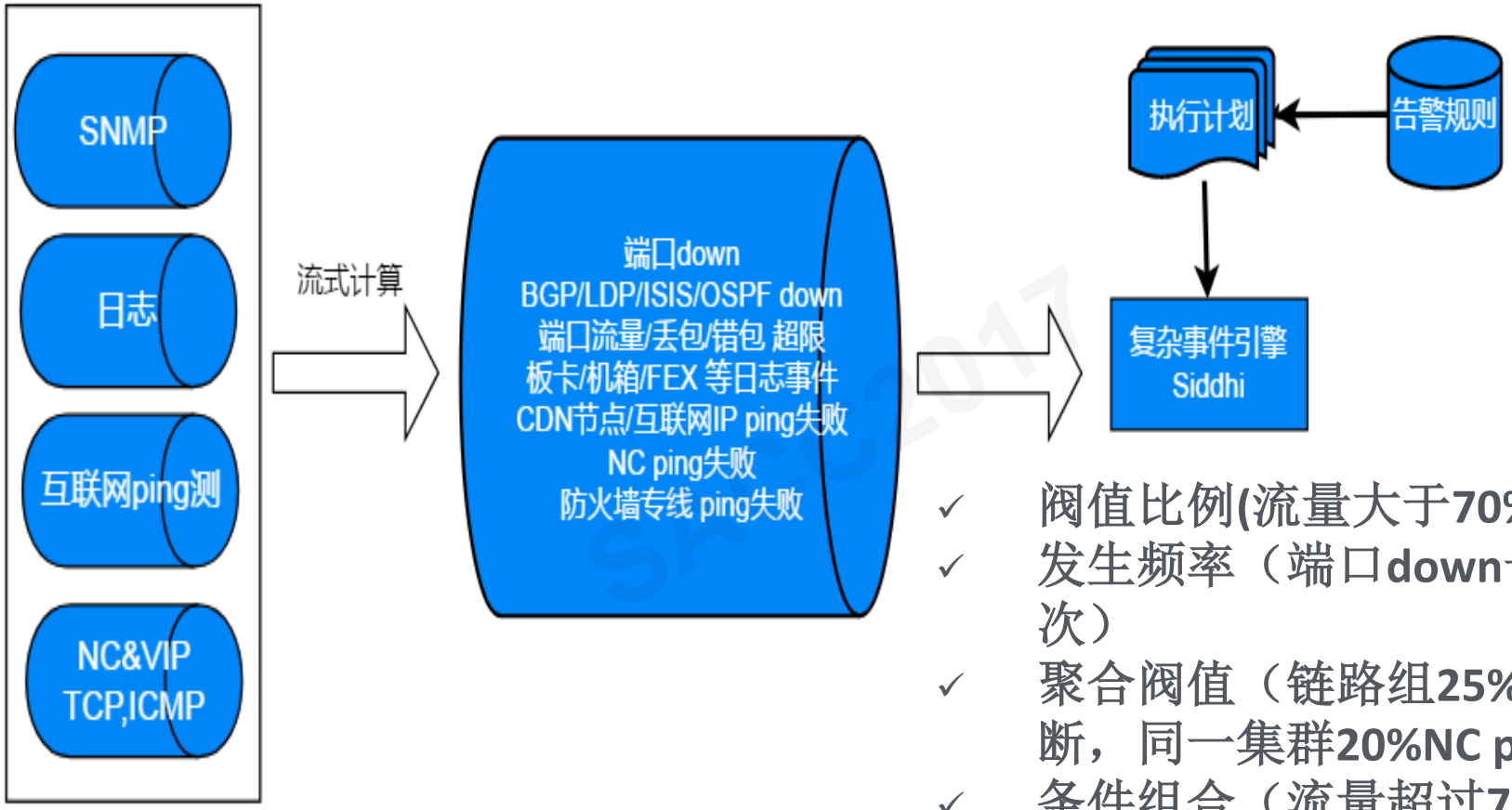
运营商切流

主备切换

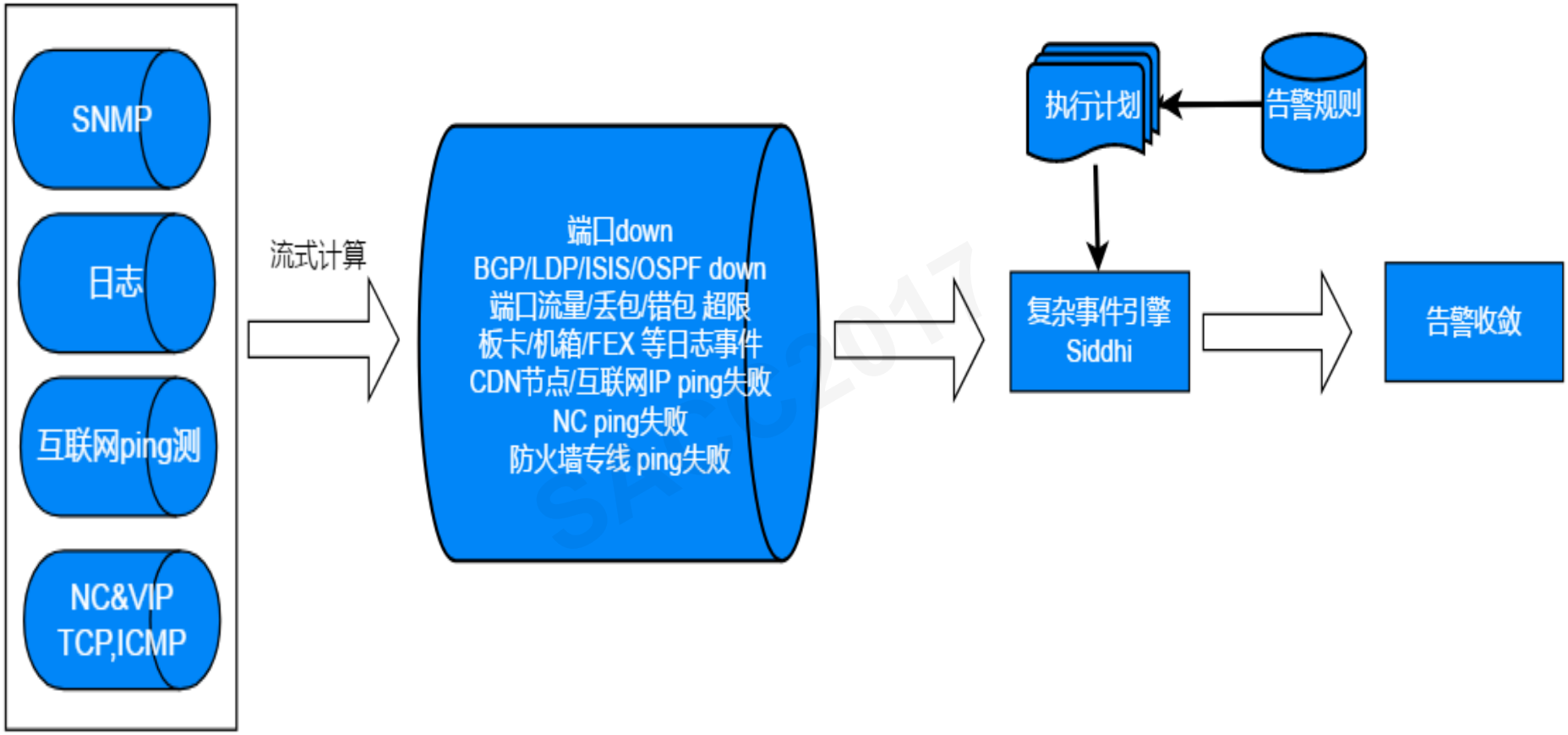
复杂场景，人工介入



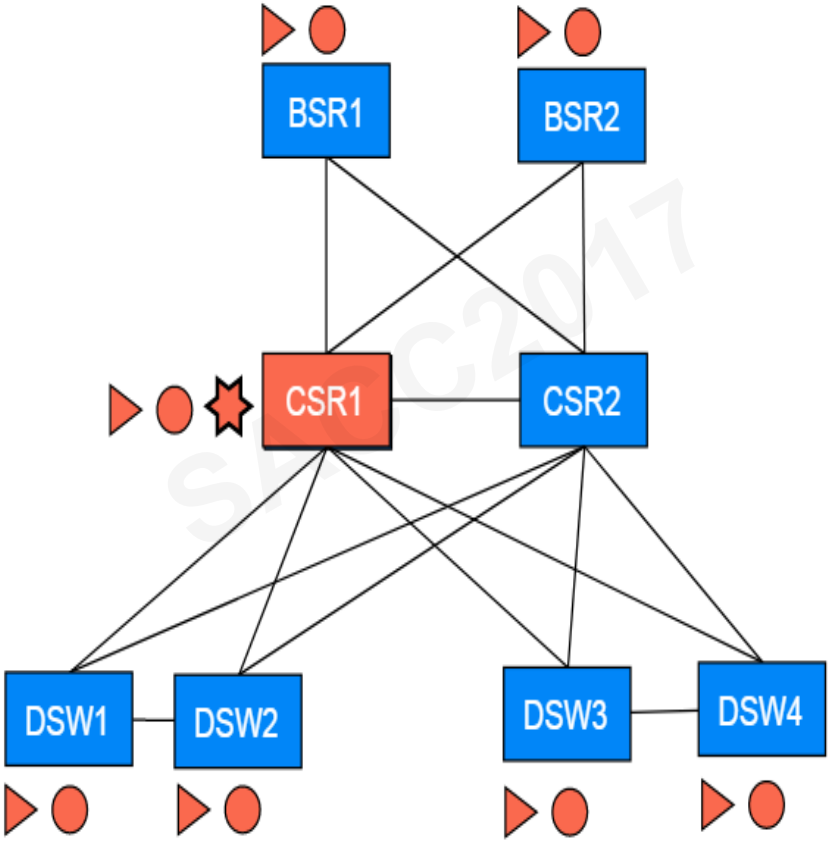

SNMP
日志
互联网ping测
NC&VIP
TCP,ICMP
流式计算
端口down
BGP/LDP/ISIS/OSPF down
端口流量/丢包/错包 超限
板卡/机箱/FEX 等日志事件
CDN节点/互联网IP ping失败
NC ping失败
防火墙专线 ping失败



- ✓ 阀值比例(流量大于70%)
- ✓ 发生频率（端口down一分钟十次）
- ✓ 聚合阀值（链路组25%链路中断，同一集群20%NC ping失败）
- ✓ 条件组合（流量超过70%并且出现丢包）


SNMP
日志
互联网ping测
NC&VIP
TCP,ICMP
流式计算
端口down
BGP/LDP/ISIS/OSPF down
端口流量/丢包/错包 超限
板卡/机箱/FEX 等日志事件
CDN节点/互联网IP ping失败
NC ping失败
防火墙专线 ping失败
执行计划
告警规则
复杂事件引擎
Siddhi
告警收敛

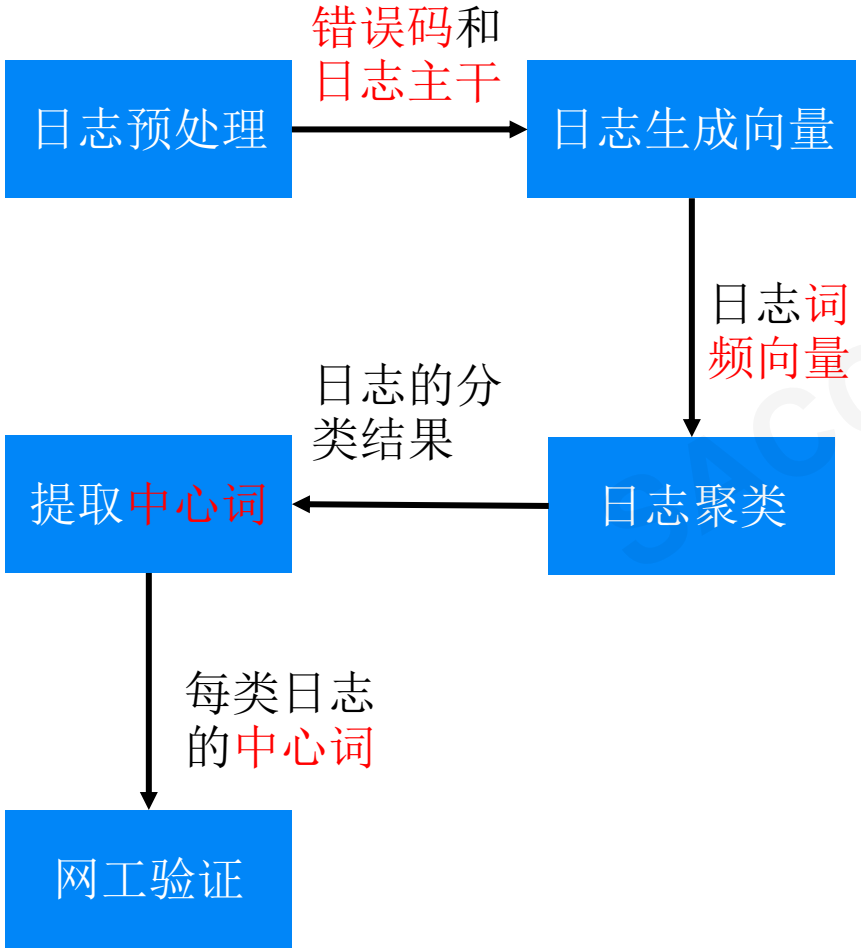
连通子图+PageRank+告警等级



告警收敛

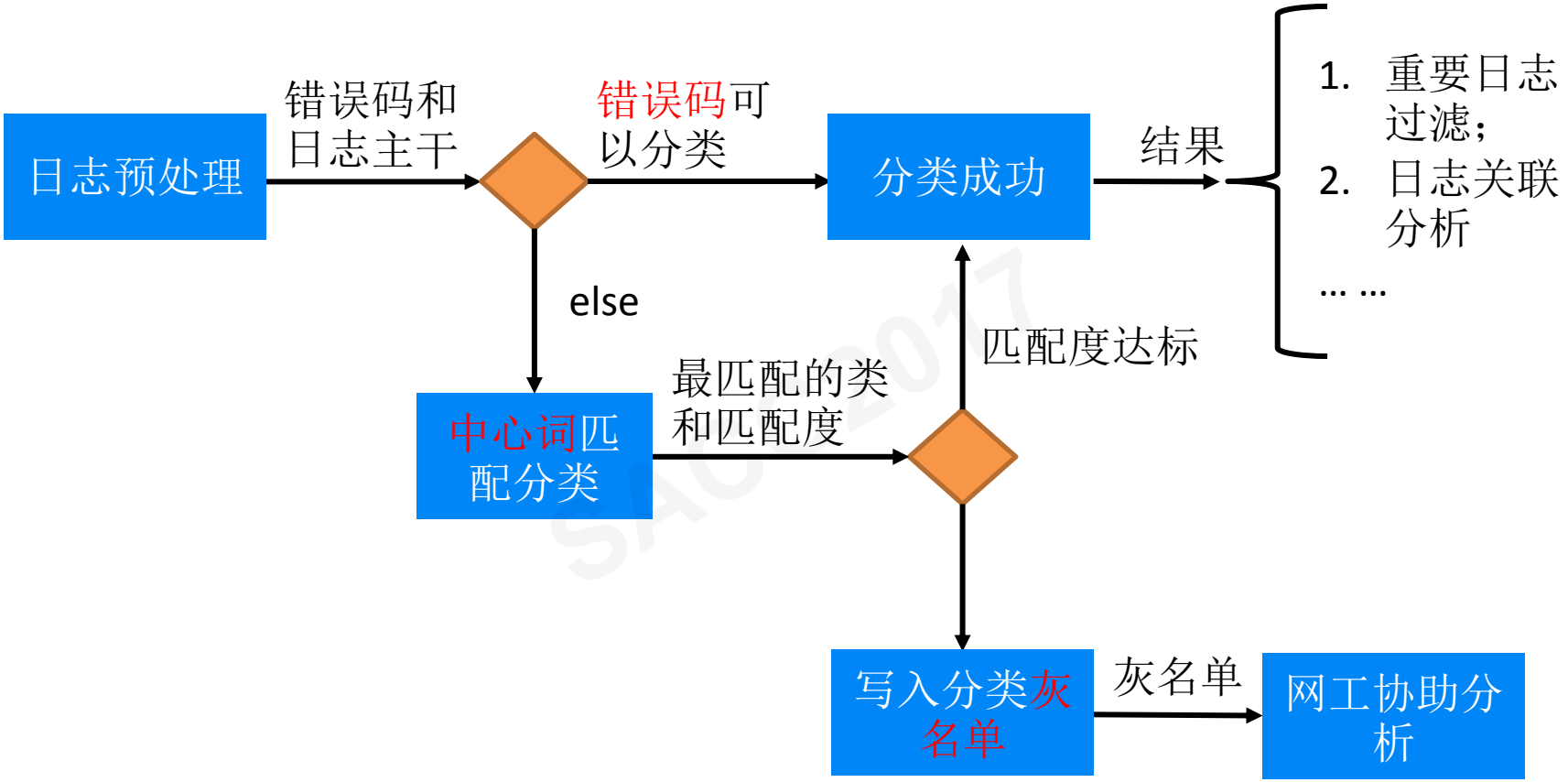
Syslog聚类分析

我们的方法：人工干预聚类结果，基于Active learning使得聚类结果逐渐变优



- 错误码：日志中标记错误类型的编码。例如：PLATFORM-SFP-2-LOW_RX_POWER_ALARM
- 日志主干：
  - 将ip替换成ipaddress
  - 将mac地址换成macaddress
  - 将端口替换成phyport
  - 将数字，符号等去掉
- 中心词：可以代表一类日志的词序列。

Syslog实时分析系统



采集

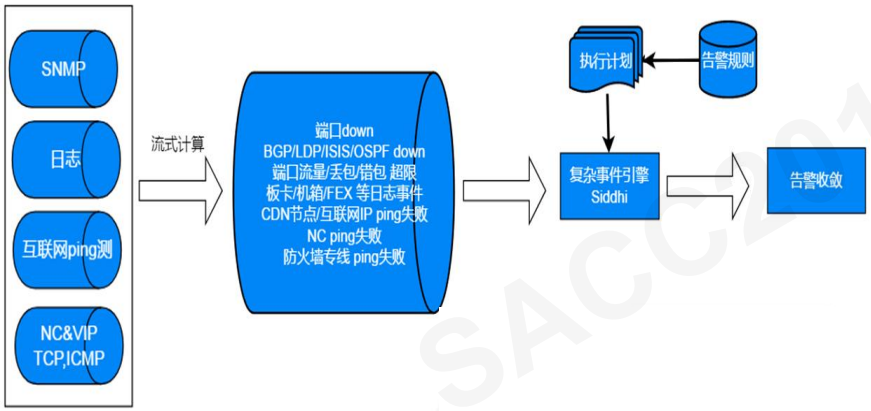
网络设备/防火墙/ANAT/LVS

互联网质量

专线&VIP

服务器丢包&延时

故障发现



根因定位

设备状态扫描

变更扫描

流量攻击扫描

运营商状态扫描

自动化故障处理

端口/板卡/整机隔离

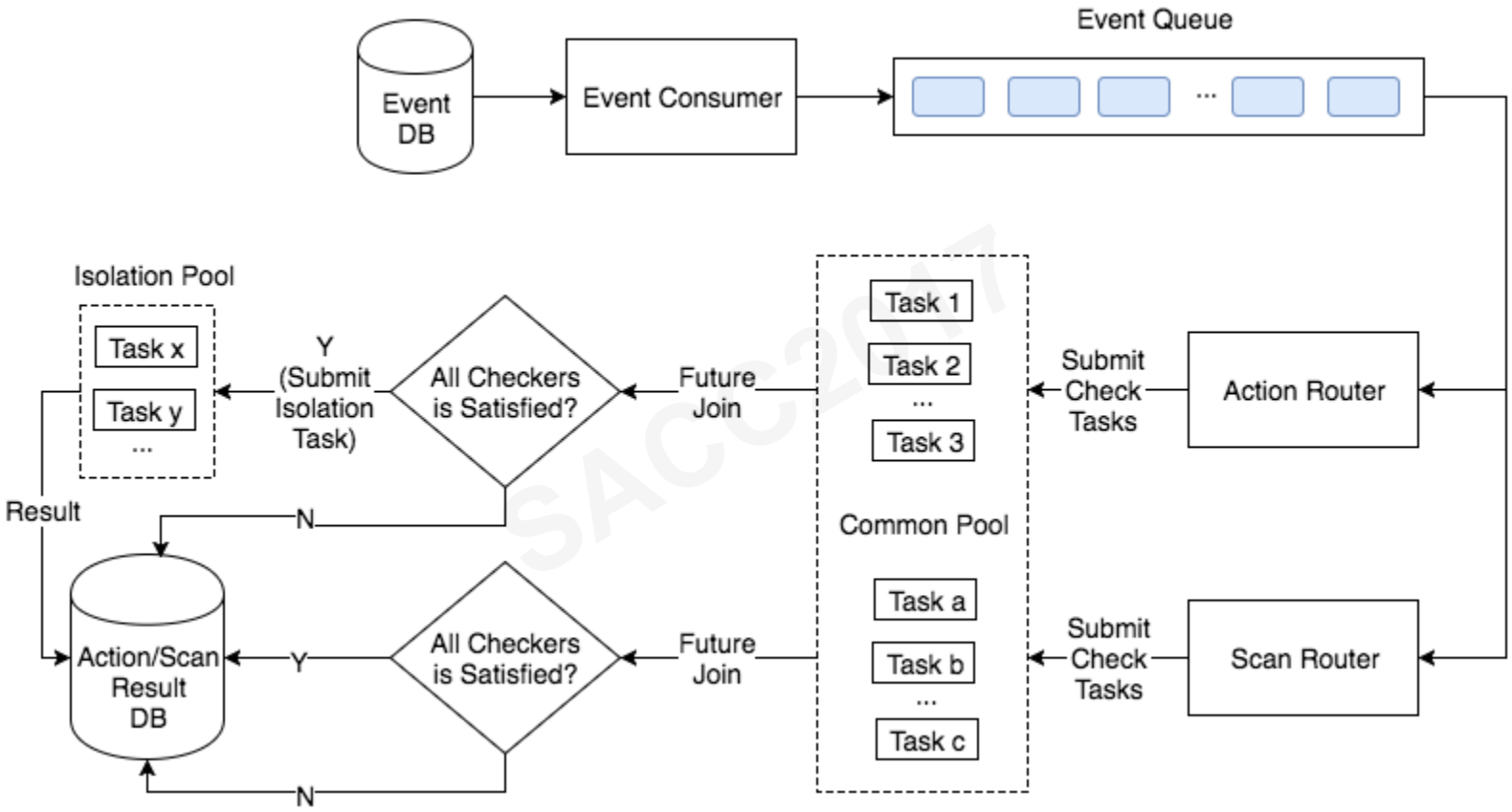
运营商切流

主备切换

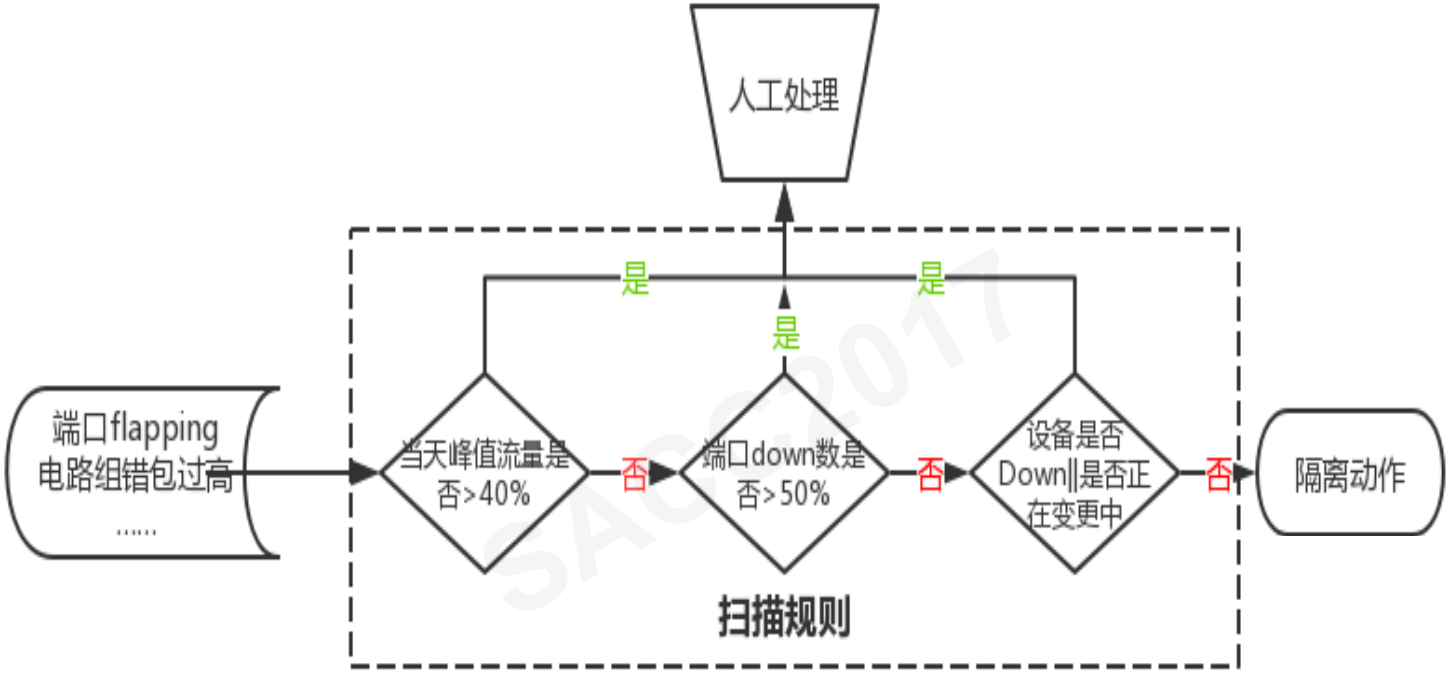
复杂场景，人工介入

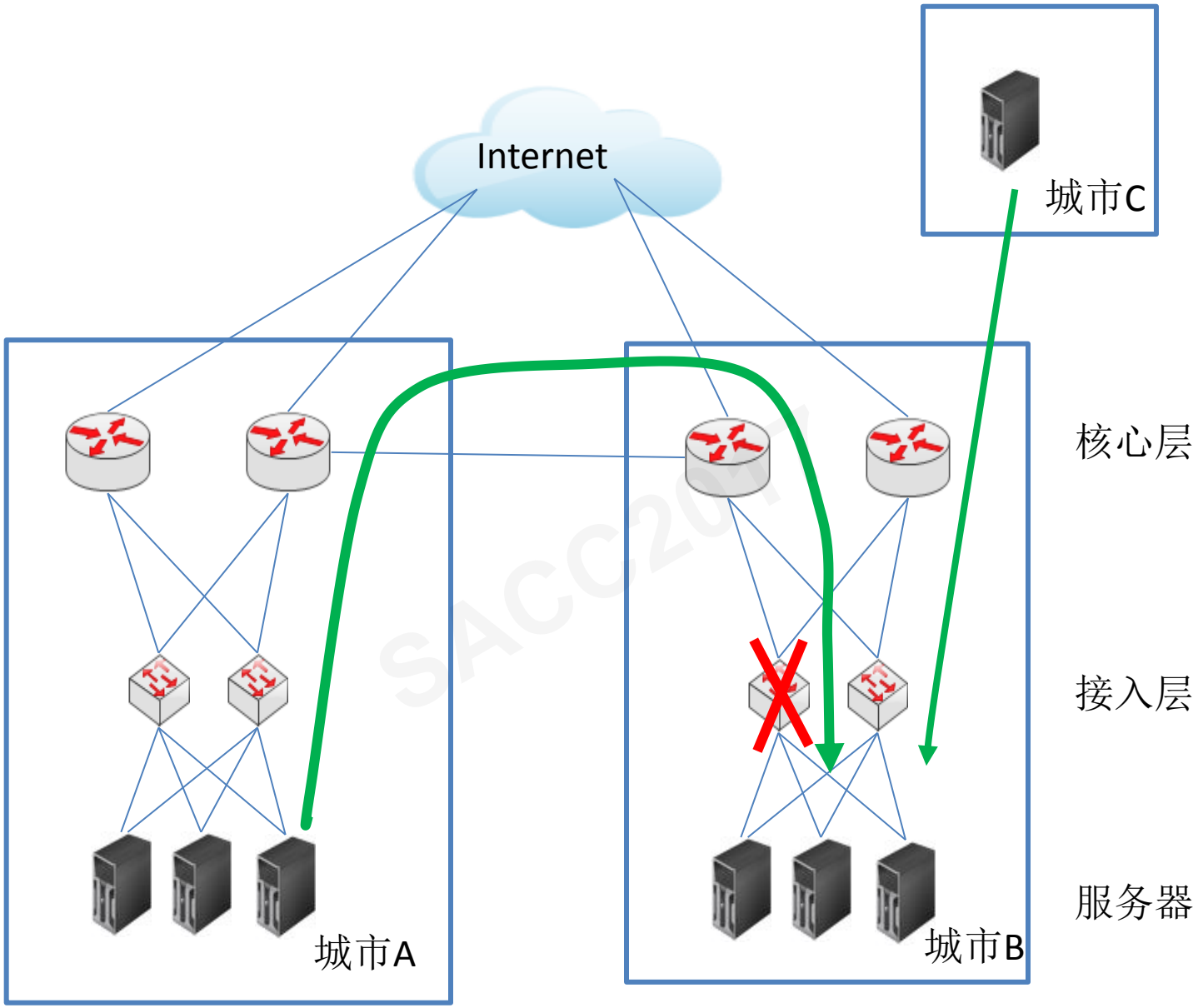
- 原子化扫描场景，告警事件触发，灵活配置

故障扫描

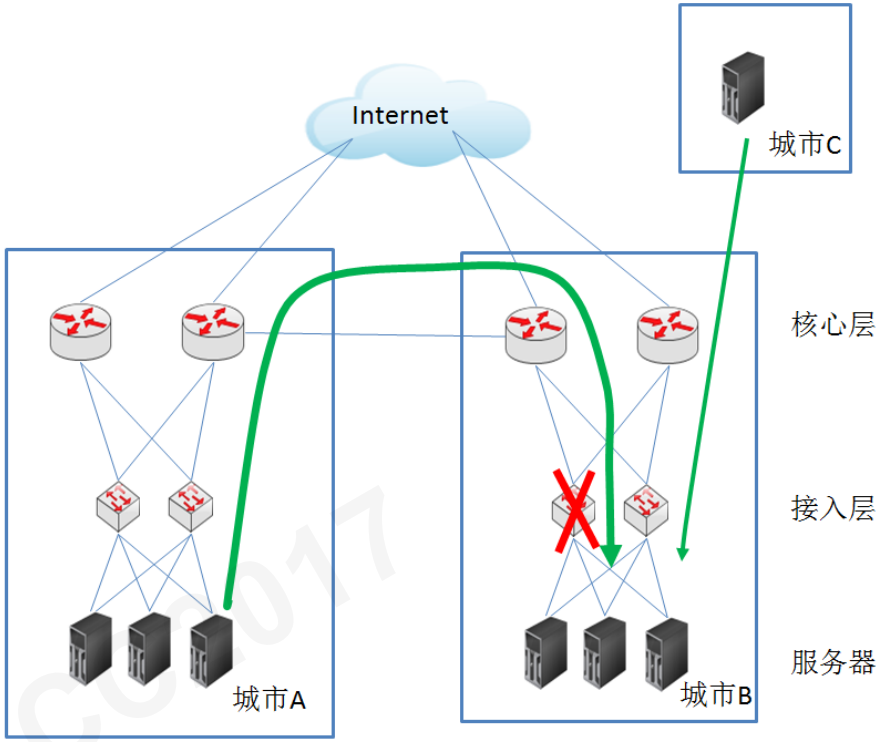


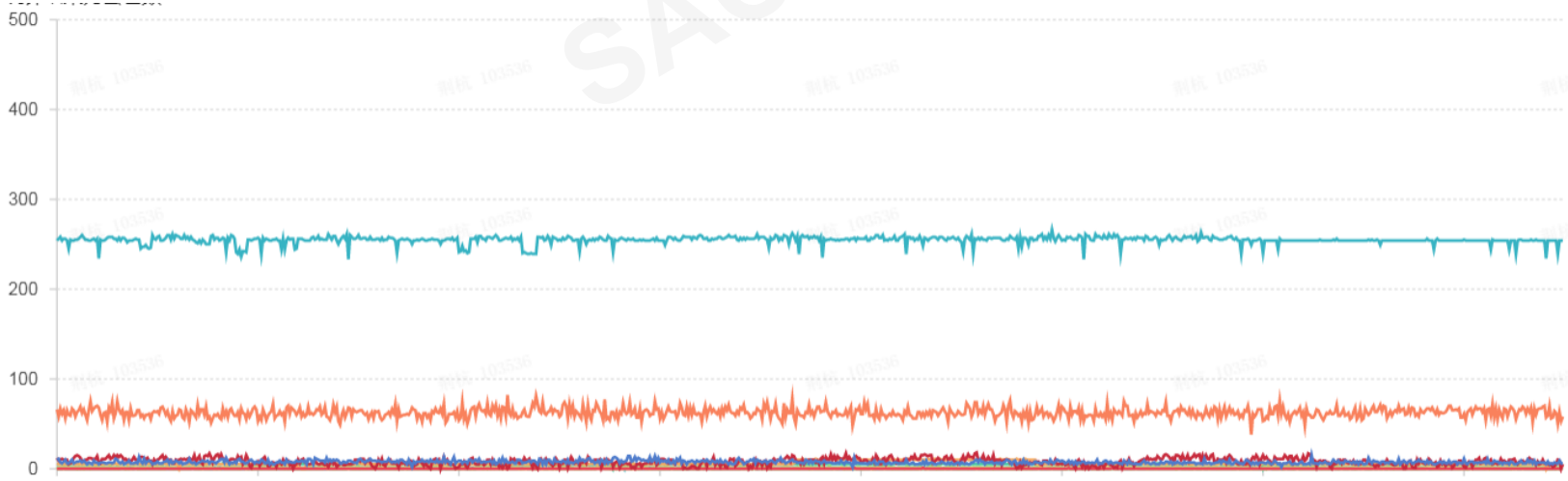
端口抖动隔离

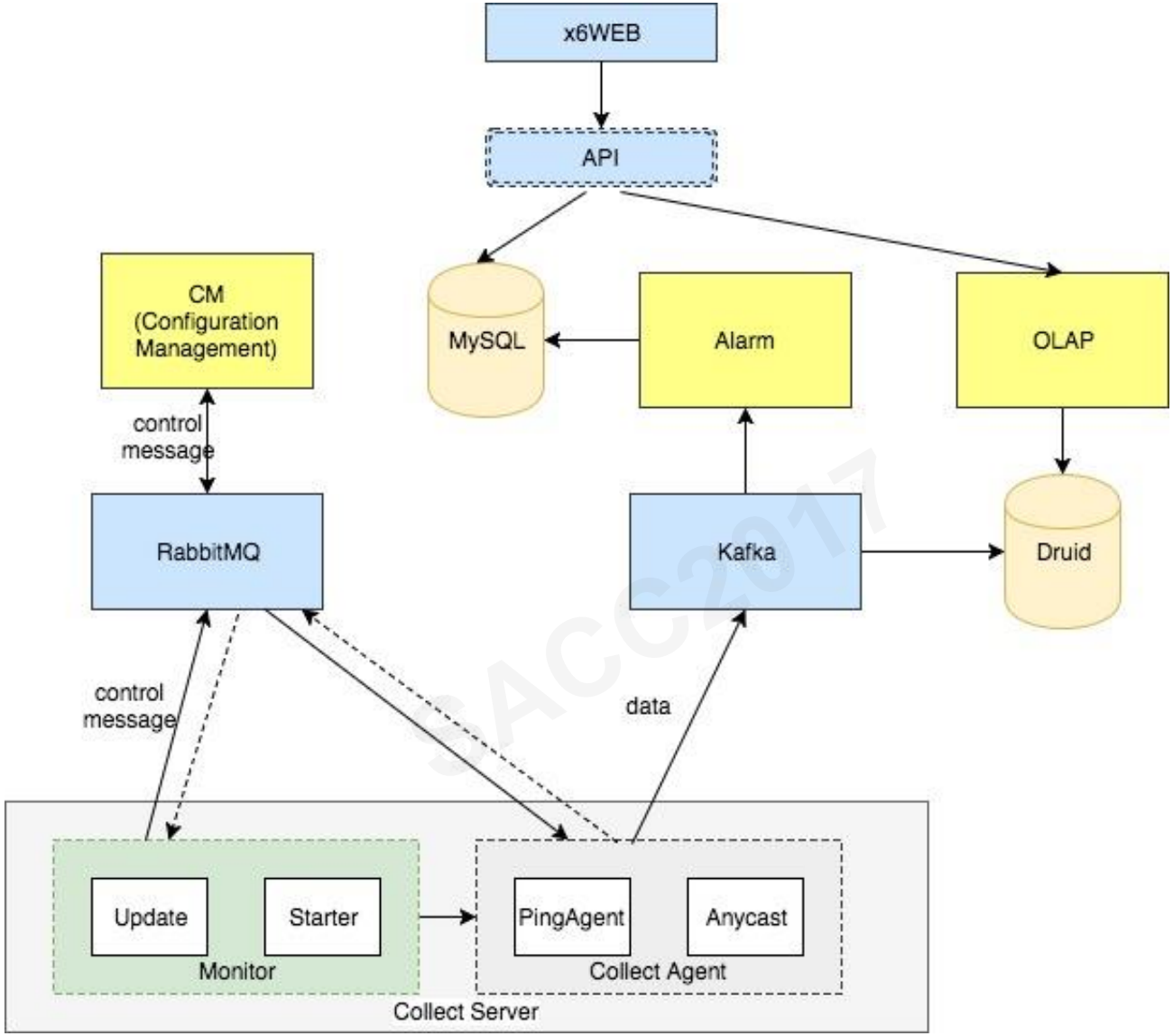




质量探测体系

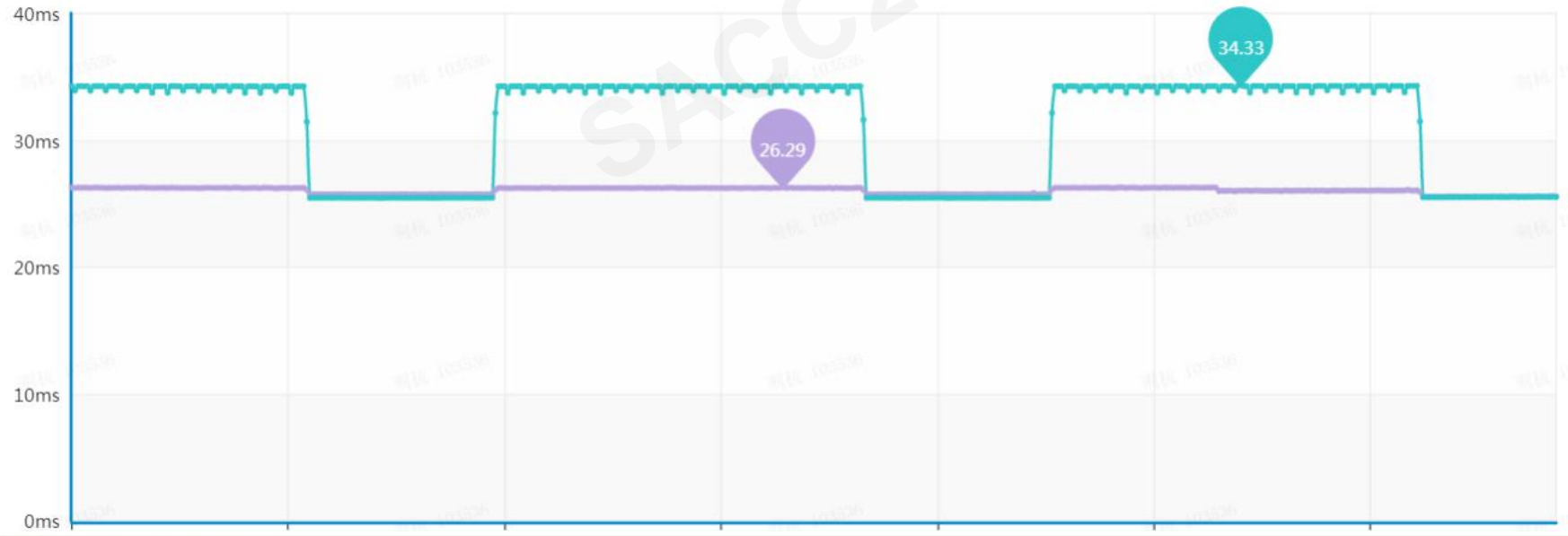





x6WEB
API
CM
(Configuration
Management)
MySQL
Alarm
OLAP
control
message
RabbitMQ
Kafka
Druid
control
message
data
Update
Starter
Monitor
PingAgent
Anycast
Collect Agent
Collect Server

互联网故障检测

- 从全球IP地址库为每个国家（中国美国的每个省和州），每个运营商动态挑选5000个存活IP进行探测，每分钟千万级IP
- 构建网络质量基线，而非单纯的阀值进行告警




THANKS