GOPS2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

会议时间：3月25日-3月26日

会议地点：深圳 • 南山区 圣淘沙酒店(翡翠店)

主办单位：开放运维联盟 OOPSA Open OPS Alliance　高效运维社区 GreatOPS Community

指导单位：数据中心联盟 Data Center Alliance

协办单位：中国新一代IT产业推进联盟

GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

由点及面，腾讯智能监控实践与思考

梁定安，腾讯

个人简介

- 梁定安（大梁）
- 10年互联网运维
- 腾讯社交平台运维负责人

目录

监控的意义和目标

- 可靠性
- 可用性
- 用户体验

监控的手段

提前埋点，应用主动上报自身的运行情况

无需埋点，从应用外部探测或获取服务的运行情况

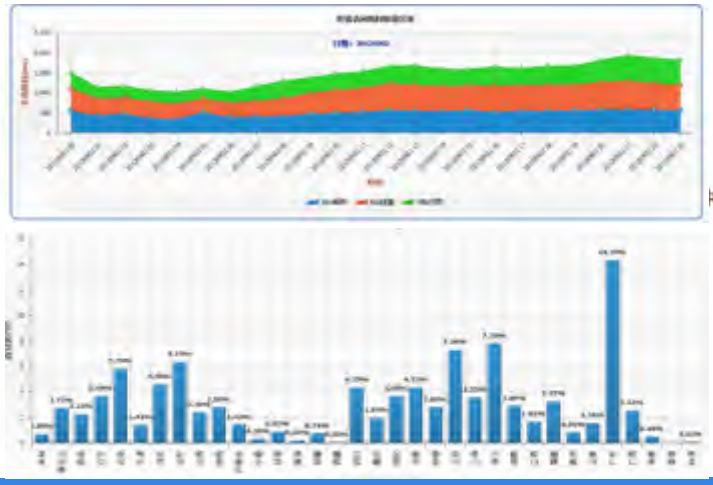
与程序逻辑无关，对服务质量的监控

监控的本质

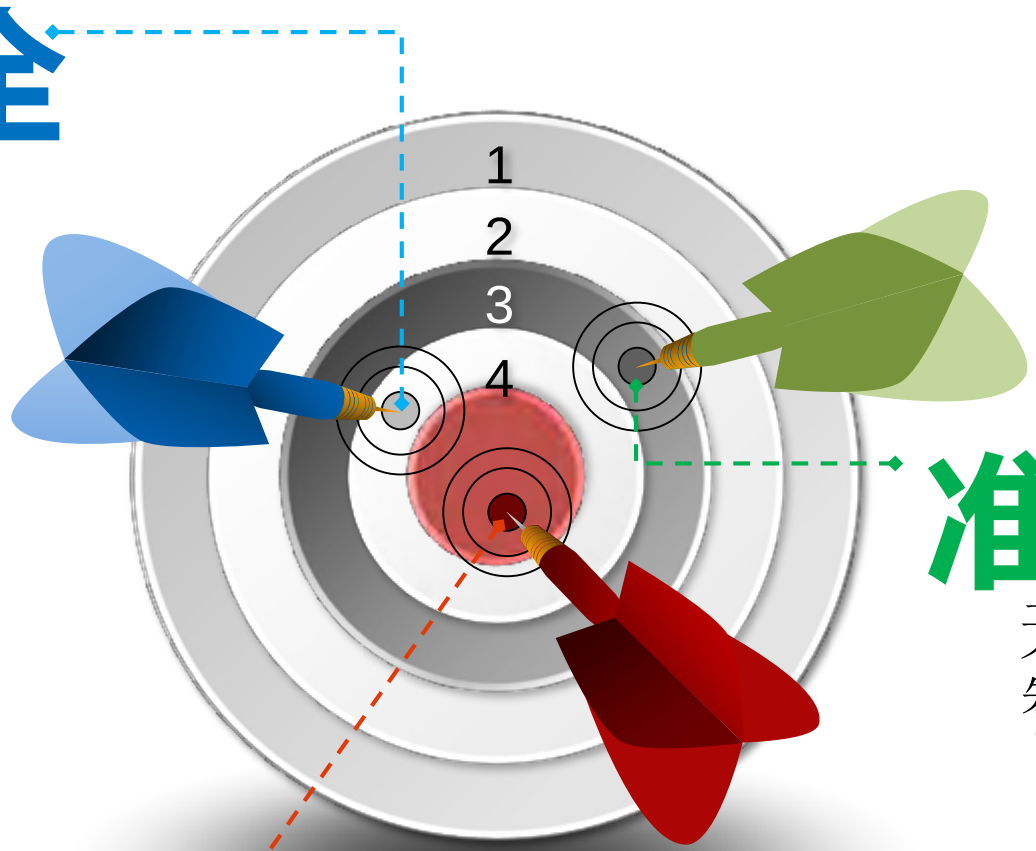
请求量
成功率
耗时

趋势
对比
波动
阀值
分布
聚类
区间

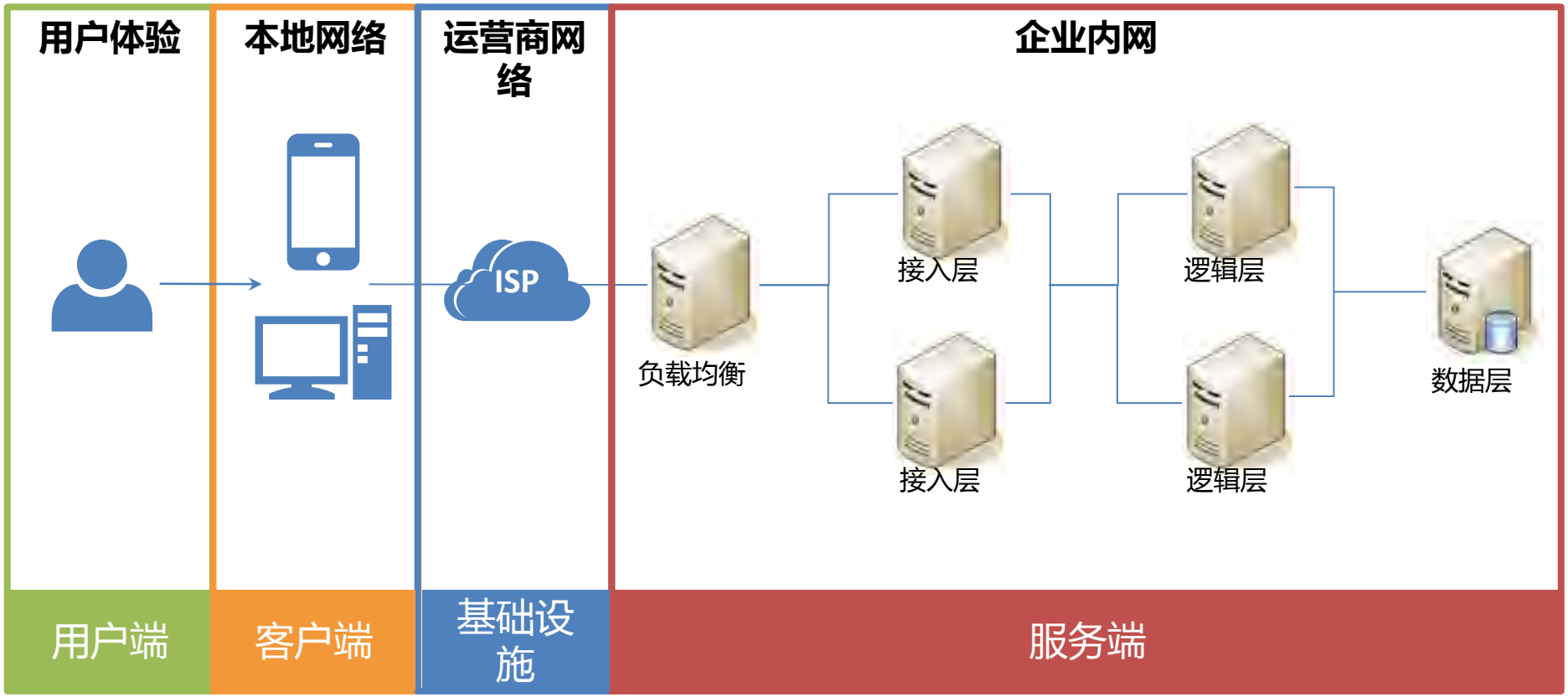
图
表
告警



监控的目标



全链路监控


用户体验
本地网络
运营商网络
企业内网
ISP
负载均衡
接入层
接入层
逻辑层
逻辑层
数据层
用户端
客户端
基础设施
服务端

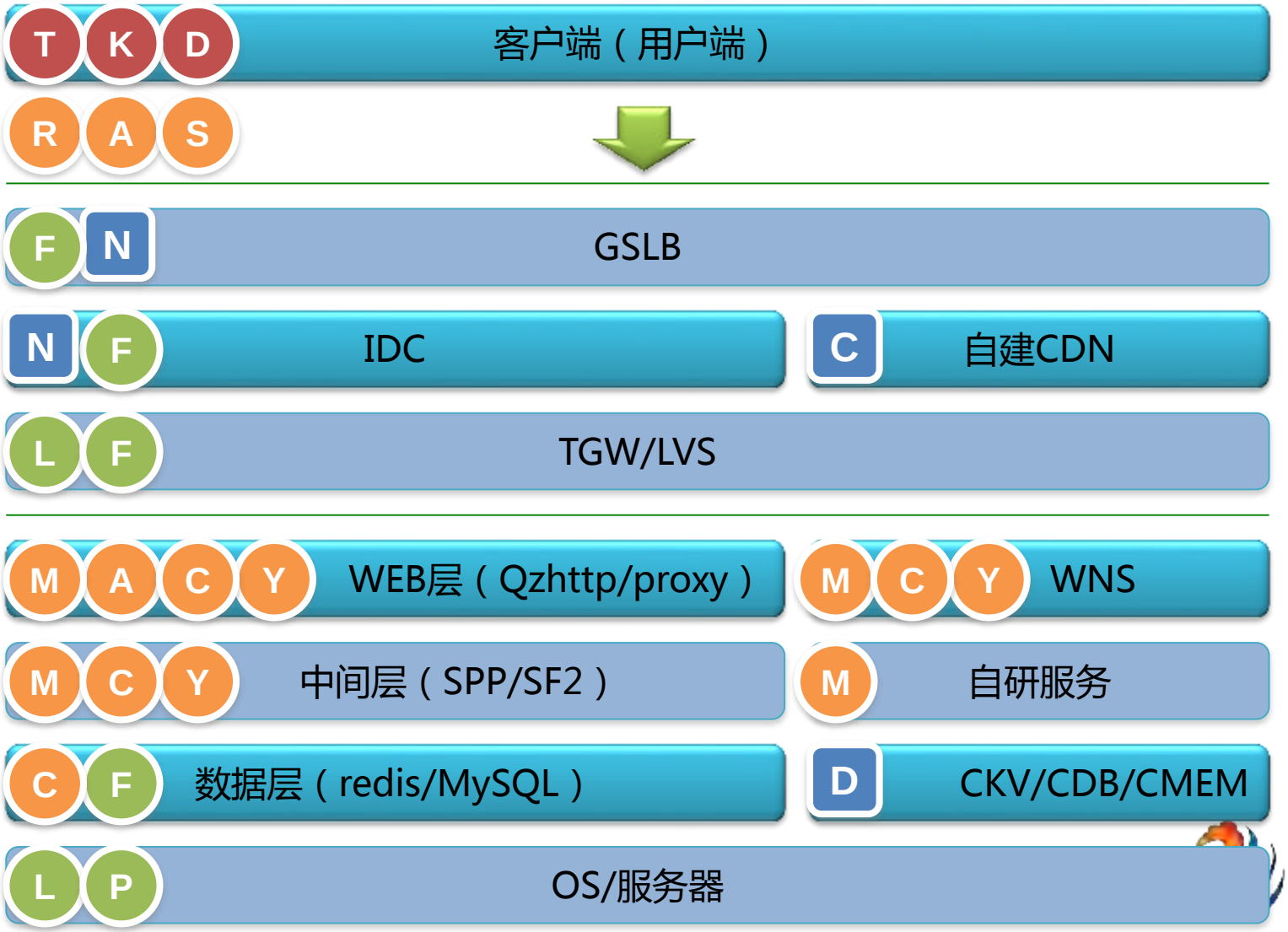
SNG监控全景图

TEG服务监控：
N: 网络质量监控
C: CDN监控
D: 数据层监控

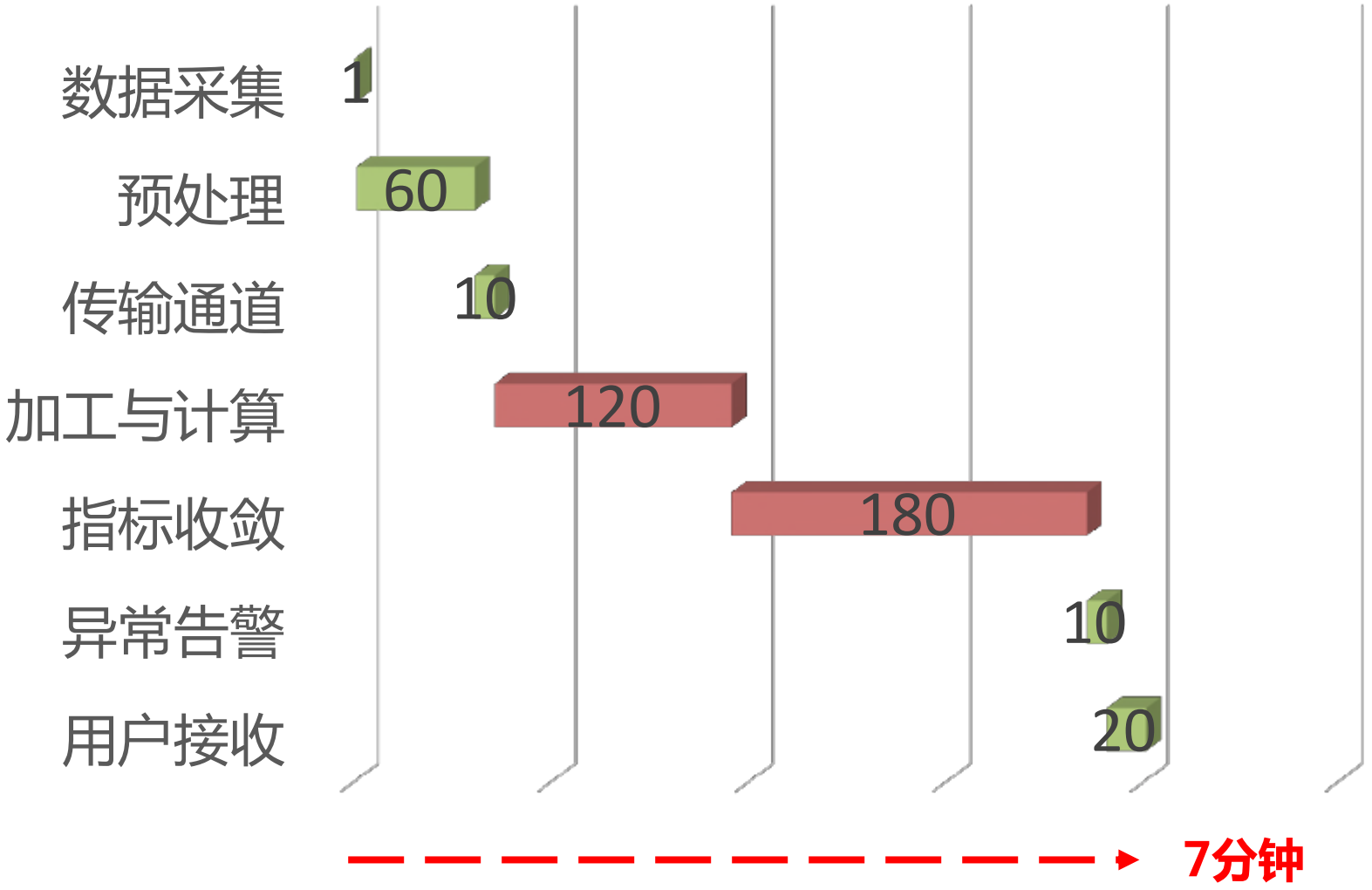
SNG服务监控：
Y: 业务染色监控
R: 返回码监控
S: 测速系统
A: 自动化测试
M: 模块间调用
C: 组件监控

基础监控：
L: 容量管理
P: 进程监控
F: 特性监控

移动端监控：
T: 舆情监控
K: 卡慢监控
D: 多维监控



监控的速度



统一上报协议

三维

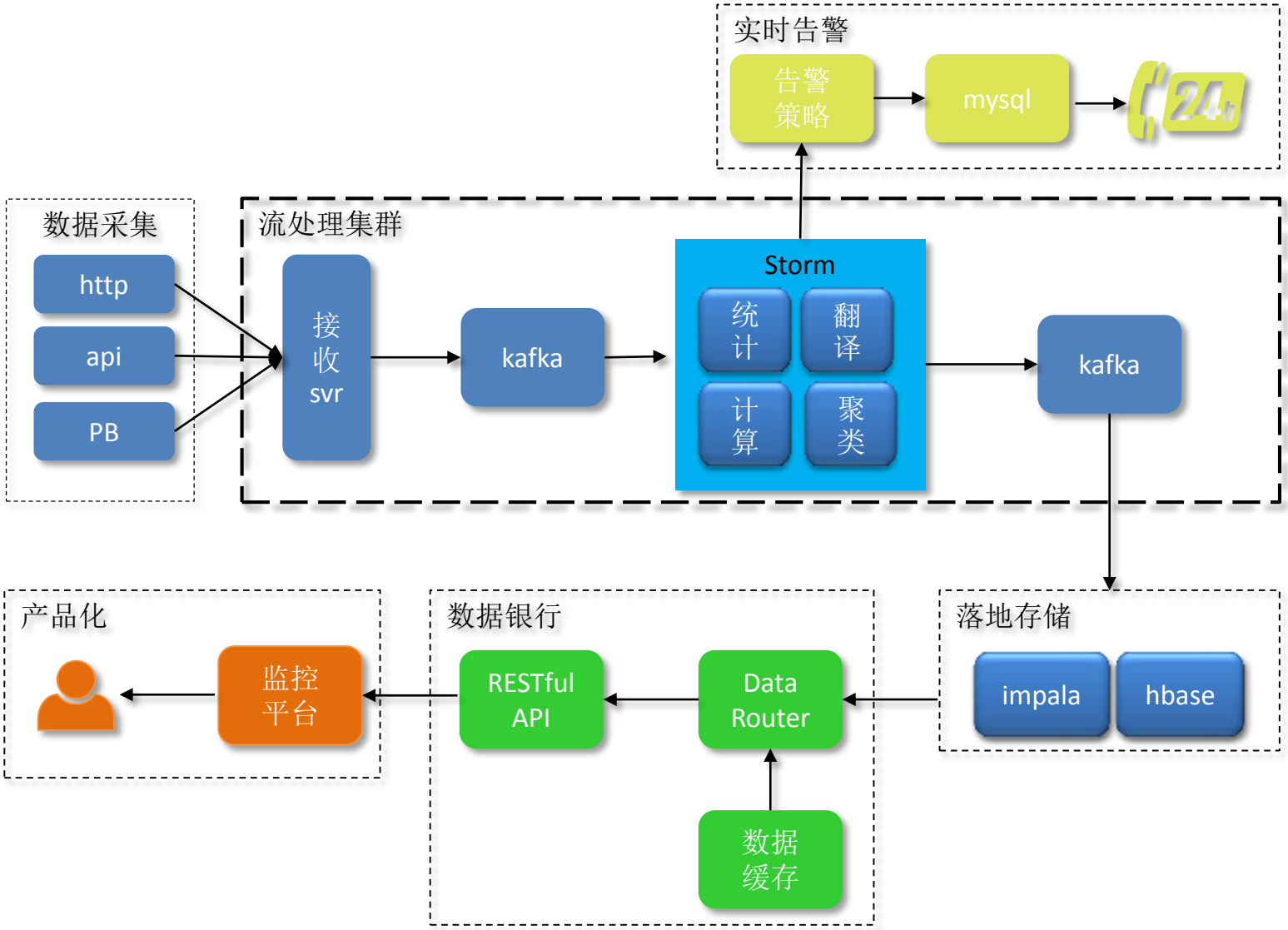
ID，IP，时间，值 → **按ID/IP/时间聚合统计**

多维

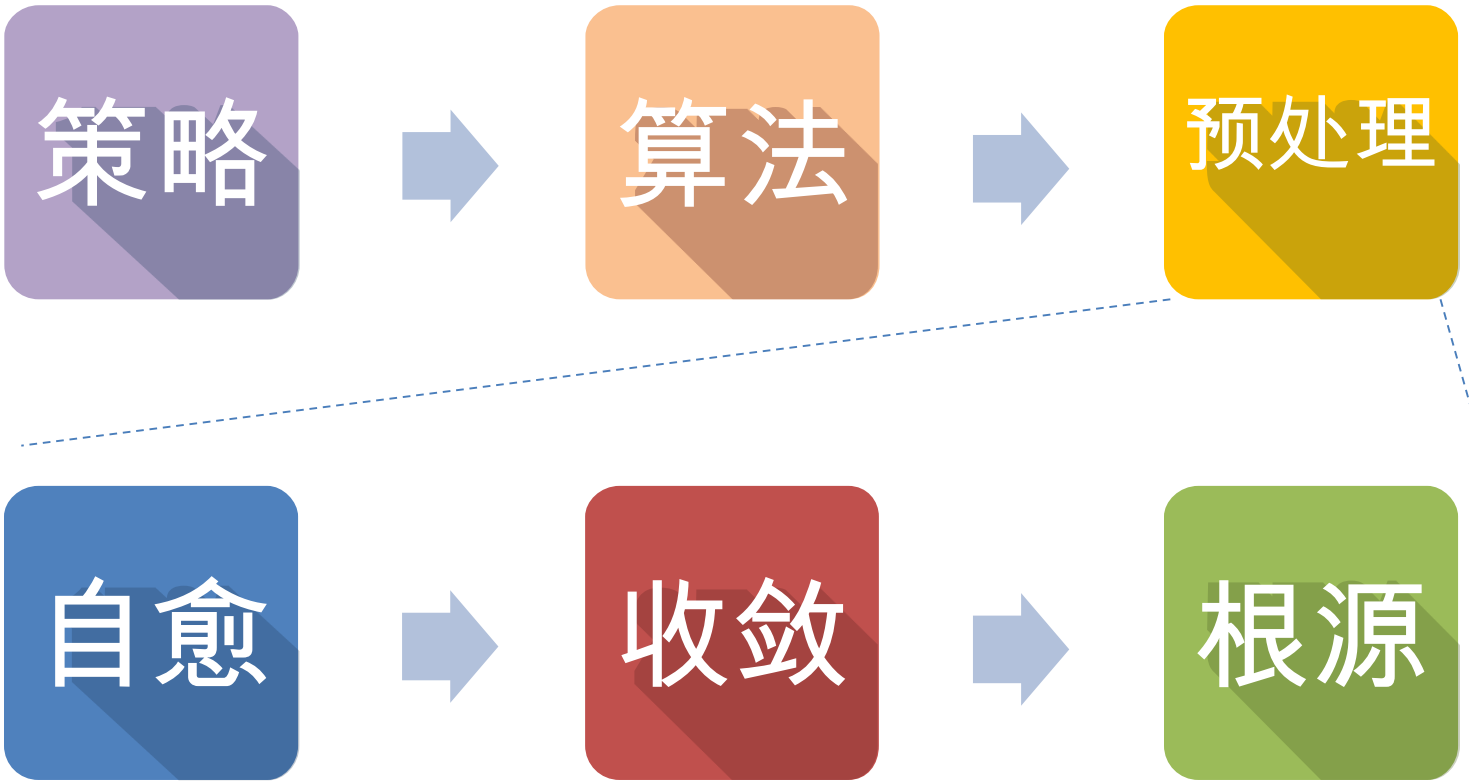
ID，时间，运营商，版本号，QQ号，业务类型，播放状态，播放页面url，请求段播放时长，完整播放时长，视频文件下载地址，播放id，请求id，用户下载速度，播放器版本号，命令字……

→ **按场景分类多维度组合**

监控集群能力


实时告警
告警策略
mysql
24h
数据采集
http
api
PB
流处理集群
接收svr
kafka
Storm
统计
翻译
计算
聚类
kafka
产品化
监控平台
数据银行
RESTful API
Data Router
数据缓存
落地存储
impala
hbase

准：智能监控



自愈


agent
状态等数据上报
TNM2
UWORK
SNG
相关故障的
预处理
告警
处理模块
不需告警
需告警
通知
值班系统
引入
UWORK
优化告警
自动化处
理的
TNM2
①
NOC
②
BG相关工具
③
告警通知


基础监控
客户端监控
服务端监控
用户端监控

海量监控的困惑

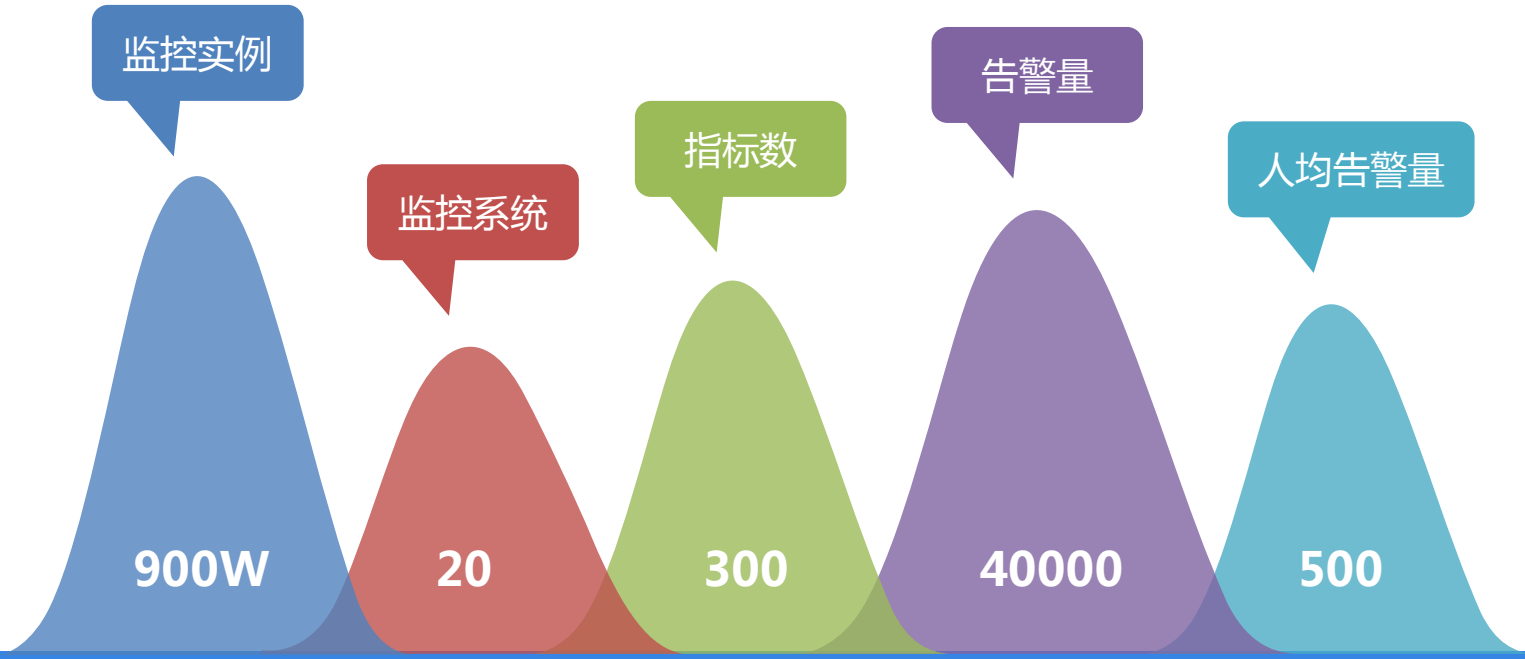
ROOT智能监控

包袱

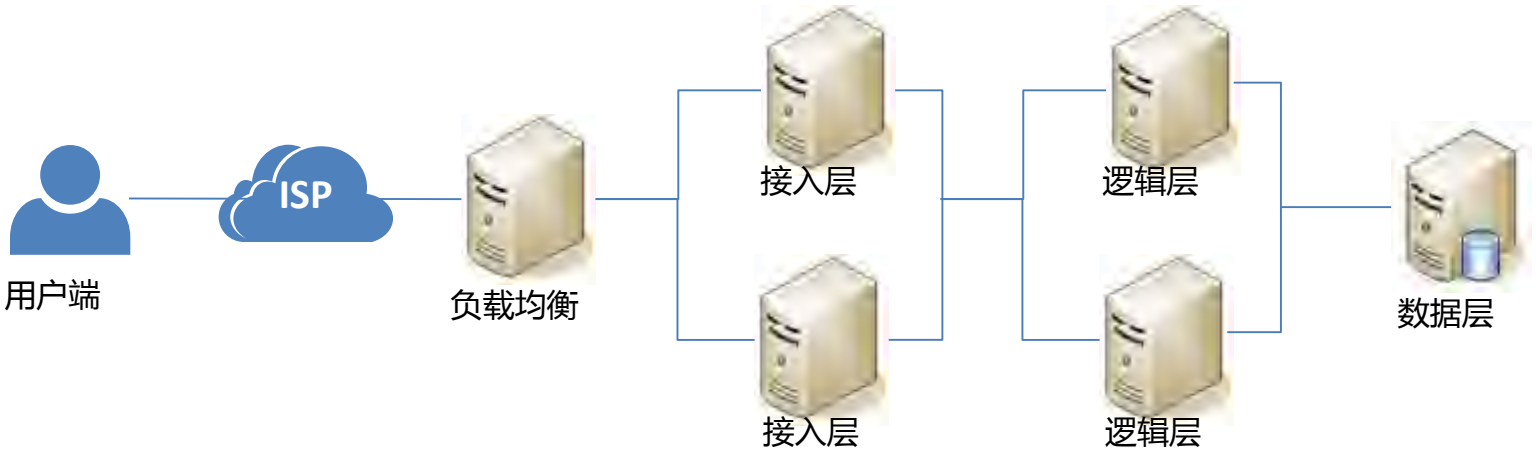
- 业务架构庞大而复杂
- 大量现象告警（点）
- 告警收敛无法最大化
- 原因告警（端到端）被淹没

ROOT

基于业务架构，结合数据流关系，通过时间相关性、面积权重等算法，将监控告警进行筛选分类，发掘有业务价值的告警，并直接分析给出告警根源



ROOT示意图



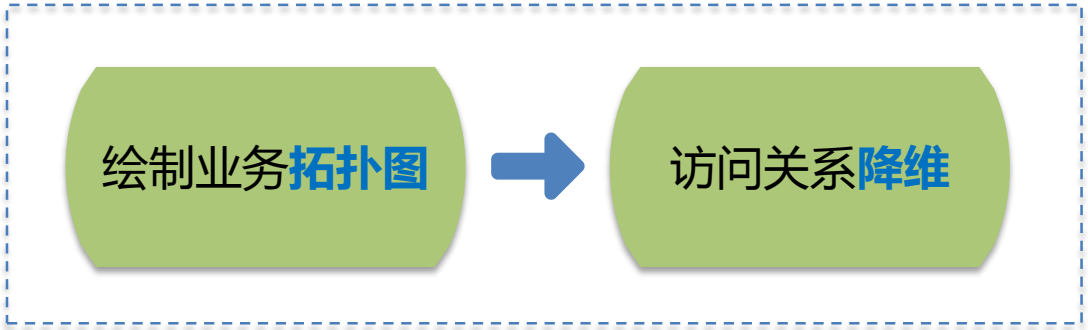
假如：DB宕机。

现实：用户端、接入层、逻辑层、数据层的监控点均有 **N** 个告警产生。

理想：智能定位到数据层监控，只发出 **1** 个告警。

ROOT分析原理

基础数据

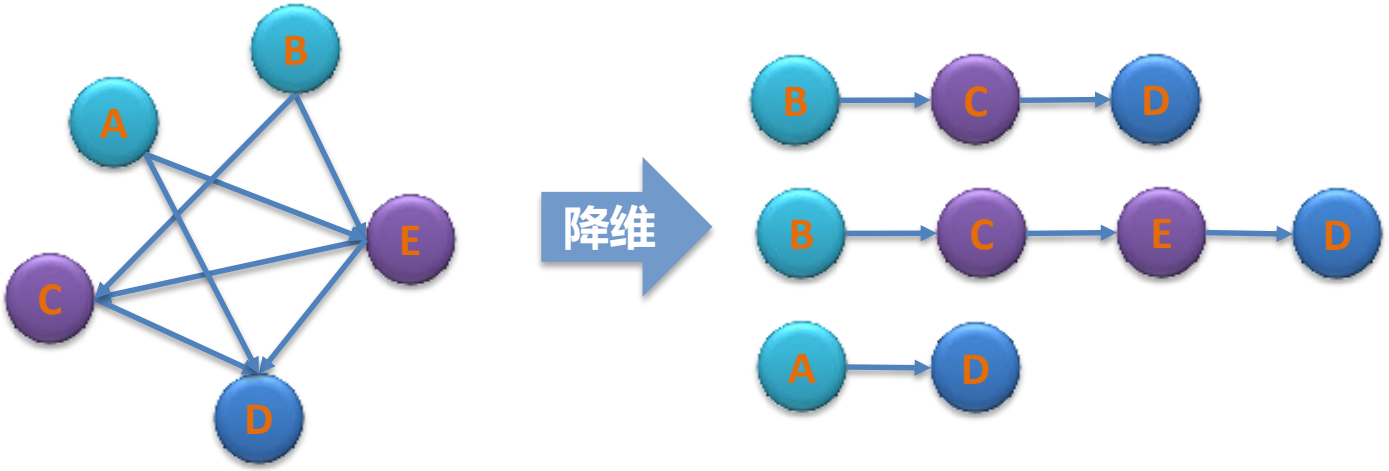


计算逻辑

降维策略

绘制业务**拓扑图** → 访问关系**降维**

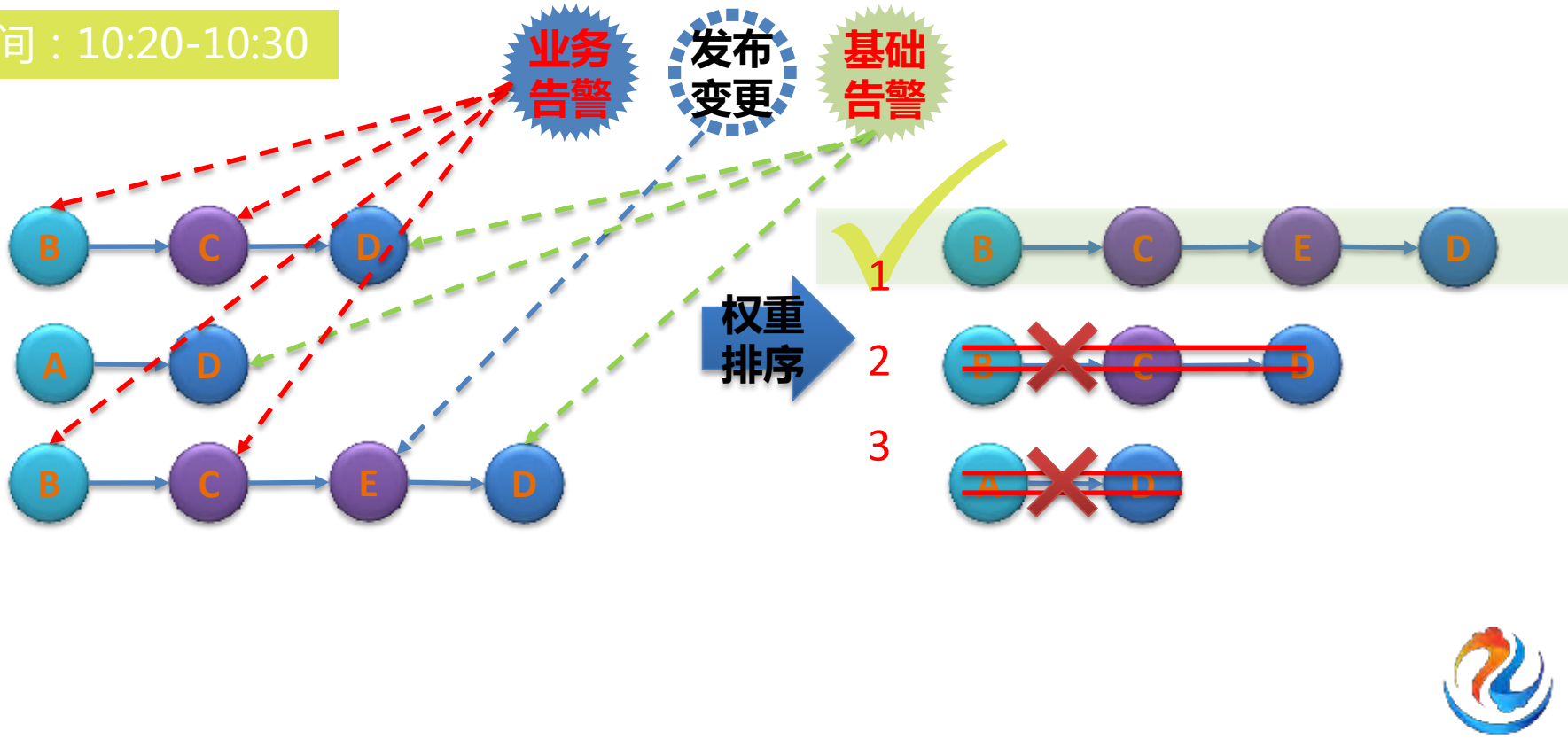
- **L5访问关系**
- **模调关系**
- **IP间抓包**



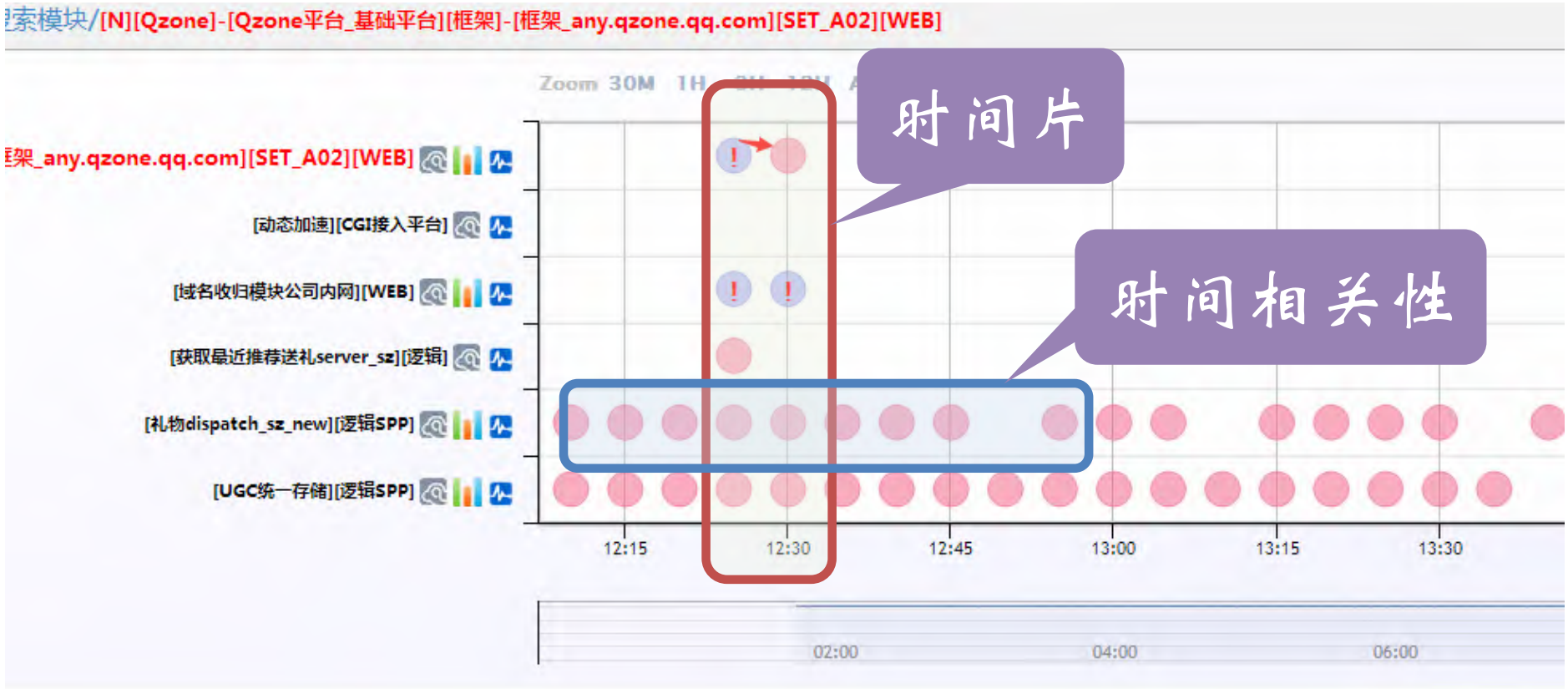
关联分析

相关时间告警**叠加** → 链路**权重**计算 → **过滤**噪点 → 原因**模型**分析 → **根源**告警

时间：10:20-10:30

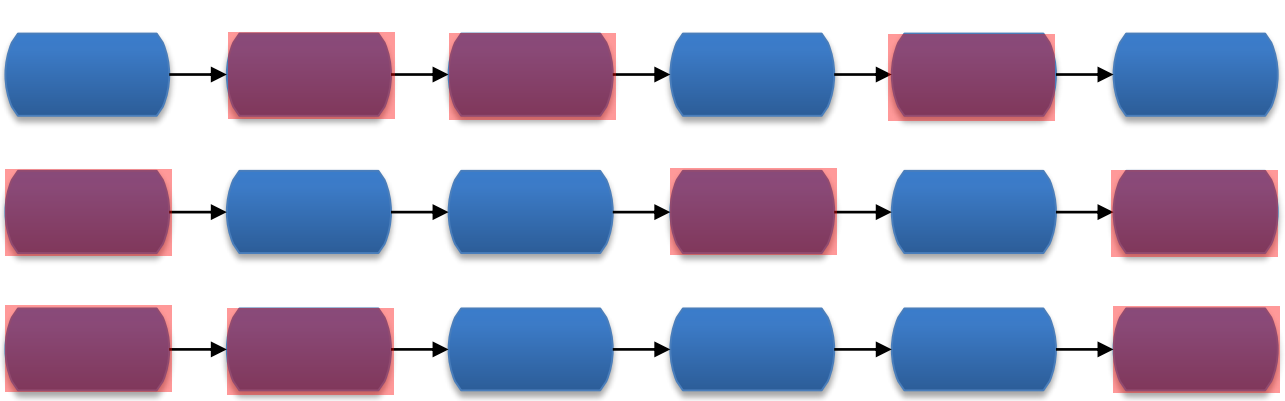


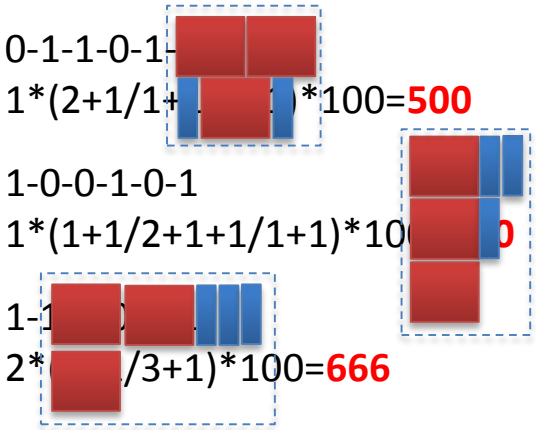
时间相关性分析



关联告警准确性：时间有效性，持续性，告警时延，链路相关性。

权重面积分析





- ✓ **链路中告警模块数=1**

  长=1（只有一个模块告警时固定为1），宽=（1+告警模块所在链路的序号除以链路总模块数），面积=长*宽=1*（1+（iarr+1）/lnkcout）*100

  a、1-0-0-0, 权重面积=1*（1+（0+1）/4）*100=125；

  b、0-1-0-0, 权重面积=1*（1+（1+1）/4）*100=150；

  c、0-0-0-1, 权重面积=1*（1+（3+1）/4）*100=200；

  备注：链路中只有一个模块告警，并且结合业务链路生成的特性，告警模块越靠后，权重面积越大；

- ✓ **链路中告警模块数>1**

  长=链路中连着告警模块的最大个数(iarrmax)，宽=连着或不连着告警模块宽都为1+1/(连着不告警的模块个数)，面积=长*宽=iarrmax*(1+1/N+...)*100

  a、1-0-0-0-1, 权重面积=1*（1+1/3+1）*100=233；

  b、1-0-0-1-0, 权重面积=1*（1+1/2+1）*100=250；

  c、1-1-0-0-1, 权重面积=2*（1+1/2+1）*100=500；

  d、1-1-0-1-0, 权重面积=2*（1+1/1+1）*100=600；

  e、1-1-1-0-1-0-0-1-1, 权重面积=3*（1+1/1+1+1/2+1）*100=1350；

- ✓ **特殊情况**

  1、链路中，前面模块都没有告警，但最后模块连着告警（相当于链路中全模块告警），权重面积*10；

  2、链路中，模块全告警，权重面积*10；

  a、0-0-0-1-1, 权重面积=（2*1*100）*10=2000；

  b、1-1-1-1-1, 权重面积=（5*1*100）*10=5000；

ROOT架构

各个渠道的告警产生

告警收集

业务访问关系链路池

- 访问关系**链路匹配**
- 时间片**相关性**
- 告警与变更**叠加**链路
- **权重**计算排序

- 持续
- 波动
- 关联

过滤筛选

原因告警
分析模型集

各模型

各模型

各模型

直接给出
告警根源

历史同类告警知

访问关系数据降维

业务访问关系

逻辑访问关系

配置管理信息

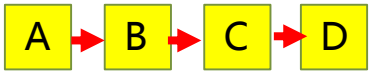

举个栗子!!

6个时间片内

告警

A：用户打开超时
B：服务调用延时高
C：组件失败率突增
D：有版本变更发布

关系链路中叠加告警信息计数



相关时间片
告警分类

C：组件失败率突增

波动

是否恢复？

寻找关联告警

A：用户打开超时
关联到
D：有版本变更发布

模型+根源+历史数据

结论
发现：[A]用户打开超时
是由于[D]有版本变更发布造成
上次同类问题发生在 上周二
责任人 王小宝

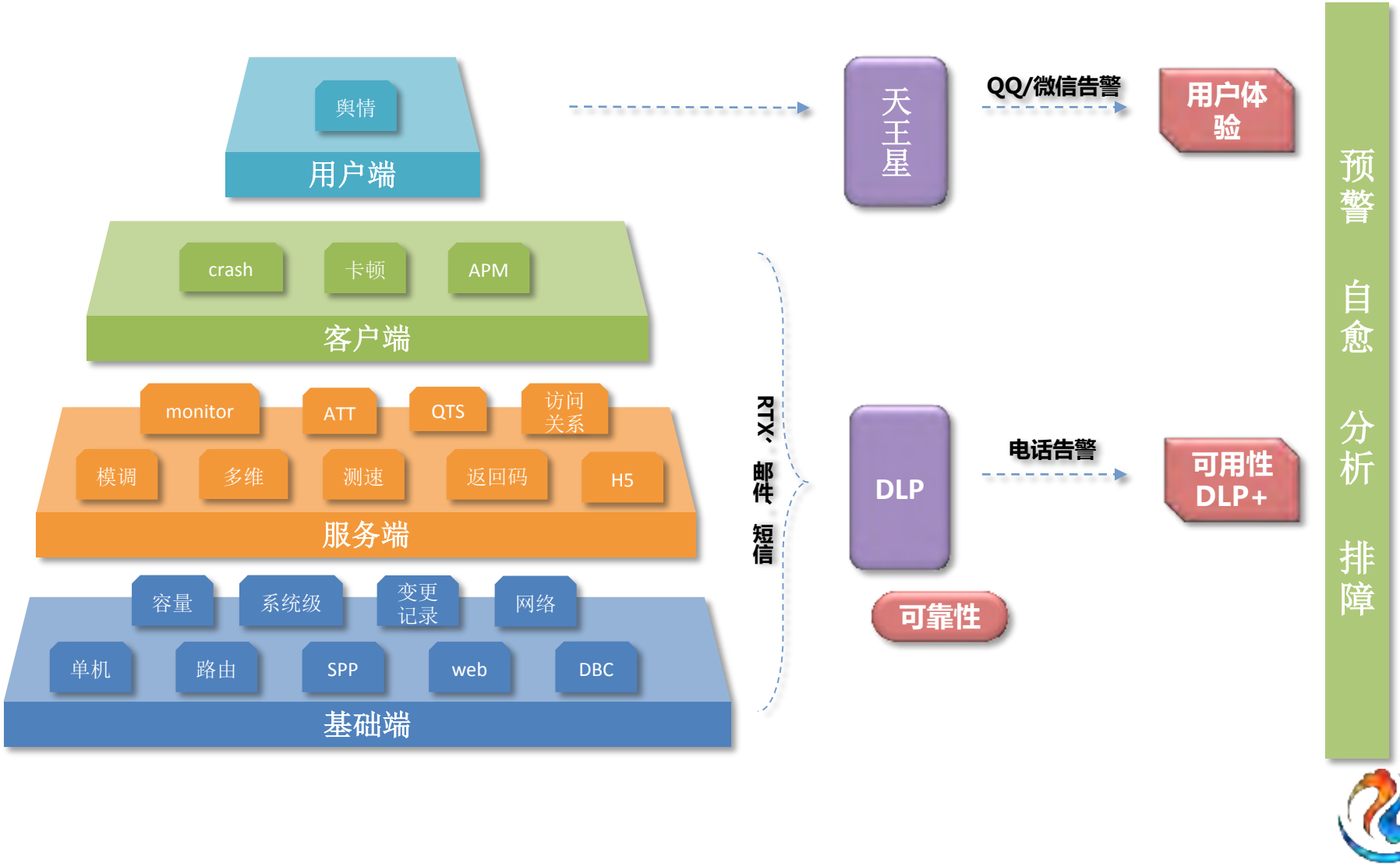
B：服务调用延时高

长期告警

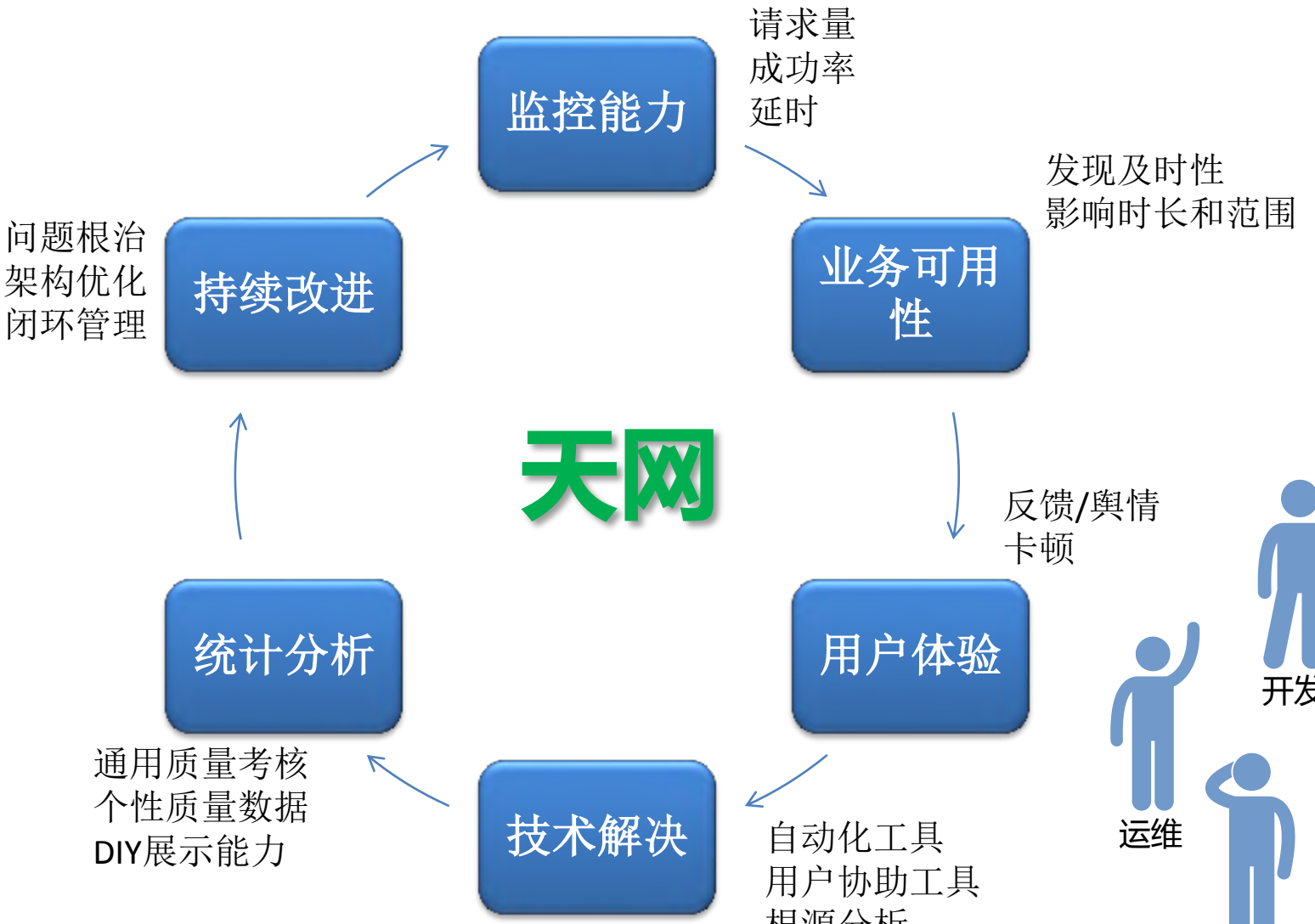
大数据分析

质量体系：生态构建

天网体系介绍


舆情
用户端
crash
卡顿
APM
客户端
monitor
ATT
QTS
访问关系
模调
多维
测速
返回码
H5
服务端
容量
系统级
变更记录
网络
单机
路由
SPP
web
DBC
基础端
RTX、邮件、短信
天王星
QQ/微信告警
用户体验
DLP
电话告警
可用性DLP+
可靠性
预警 自愈 分析 排障

天网：质量体系


监控能力
请求量
成功率
延时
业务可用性
发现及时性
影响时长和范围
用户体验
反馈/舆情
卡顿
技术解决
自动化工具
用户协助工具
根源分析
统计分析
通用质量考核
个性质量数据
DIY展示能力
持续改进
问题根治
架构优化
闭环管理
天网


开发
产品
运维
老板
客服
QA


攀登探索
GOPS 2016
Shenzhen