

# 数字化监控平台稳定性保障实践

2024年6月



<https://sre-elite.com>



**吴天昊**

**中国联通软件研究院  
副总架构师**

## 个人简介:

- 负责数字化生产运营保障体系建设与落地
- 负责数字化监控平台整体架构设计及演进
- 致力于完善“平台+应用”生态体系，打造联通集团自动化生产和智慧化运营的生产运营平台

# 目录

CONTENT

01

云原生下运维的问题挑战

02

数字化监控平台的核心能力

03

稳定性保障的场景应用实践

01

# 云原生下运维的问题挑战



## 故障如何快速发现

- 指标纷繁复杂看不全，看不清？
- 各层级数据不互通共享，铁路警察各管一段？
- 告警无人关注，处理缓慢？

## 故障如何快速定位

- 系统调用关系复杂，故障排查困难？
- 云化架构下容器服务与主机关联关系不清？
- 只知道有问题，不知道问题出现在哪里，根因无法定位？



## 故障如何快速抢通

- 需24小时运维值守，无法故障自愈及自动化？
- 故障发现无法及时拉会，故障管理质量效率低下？
- 无应急方案，应急操作时候全是问题？

## 故障如何优化预防

- 故障反复出现，复盘改进没有效果？
- 全链路性能瓶颈点和容量水位上线不知道？
- 隐患无法察觉，没有提前治理优化？



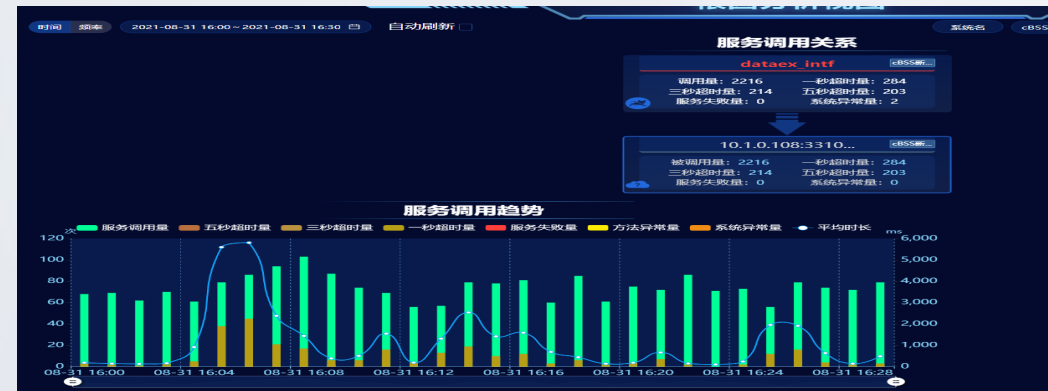
故障根因在SaaS服务下的实例



故障根因在PaaS组件



故障根因在IaaS主机



故障根因在外部接口

随着云原生技术的不断成熟，企业数字化转型也在不断加速，企业IT架构进入云原生时代，多云多集群部署已经成为常态和趋势，几何增长的云资源、微服务以及复杂化的调用关系与业务场景，传统人肉运维难以为继，如何保障系统的全面稳定，保证业务流程的高效运转，为系统运营提出了不小的挑战。

监控对象：几何级数增长，人力维护不能胜任

应用软件:

几个  
Jar包

中间件:

几套Oracle

硬件:

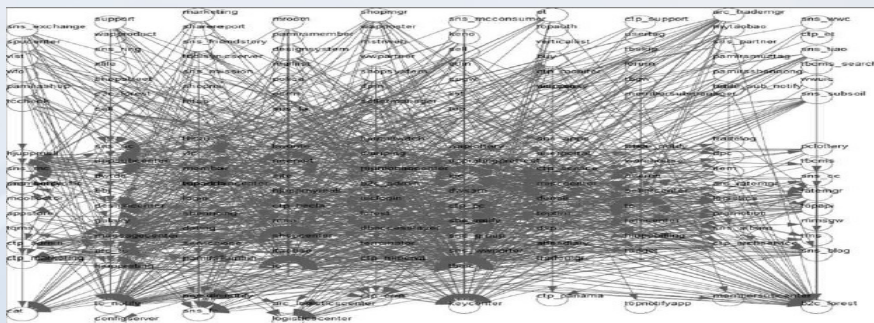
可数  
小型机

上千个  
微服务

几十种中间件清单

成千上万  
硬件

调用承载关系极其复杂，亟待引入运维工具



## 分布式架构挑战

- 维护对象：系统节点、微服务数量几何级数增加
- 调用关系：从简单对应到极其复杂，人力维护无法胜任
- 数据分片、异地存储，传统维护模式难以为继

## 运维生态挑战

- 工具重复：工具按烟囱式建设，能力分散
- 能力割裂：运维工具能力割裂不成体系
- 数据孤岛：应用、数据库、中间件、云平台、基础设施各管自身

## 业务连续性挑战

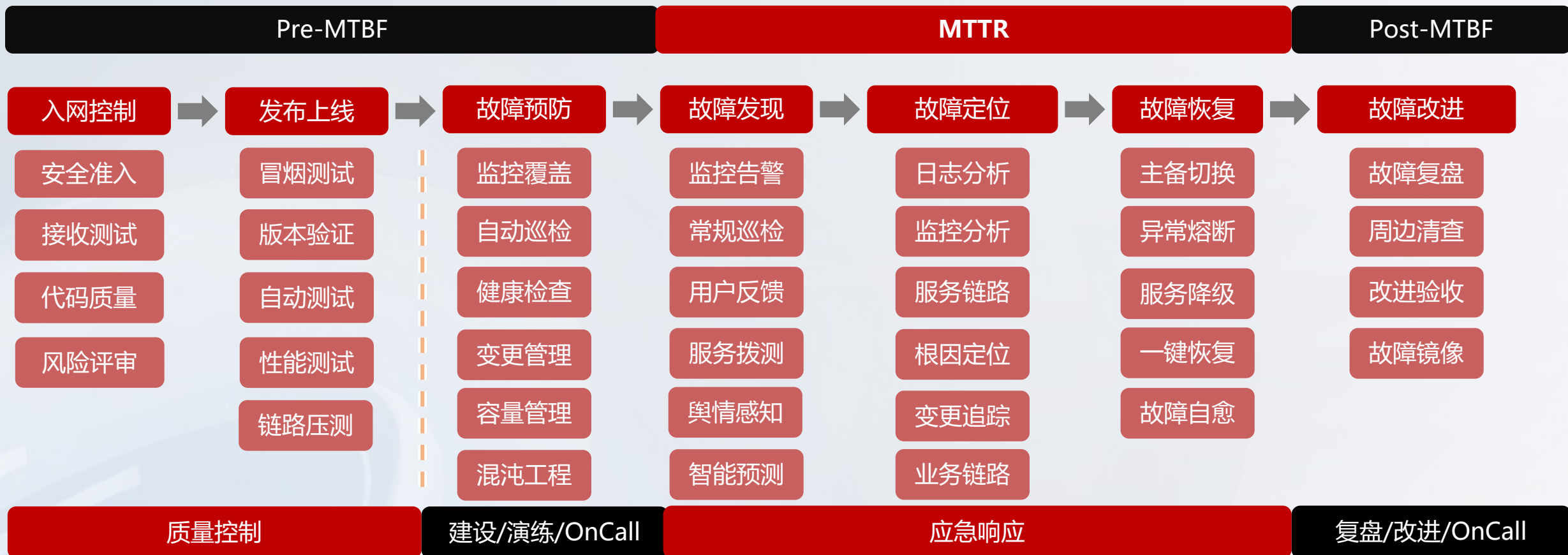
- 故障处理过多依赖专家经验，系统服务间调用链路复杂，故障分析定位困难
- 端到端的稳定性保障体系缺失，自动化、智能化故障应急处理能力不足
- 故障处于被动防御、救火，没有提前预防手段，运维大数据未被合理价值挖掘

02

## 数字化监控平台的核心能力



□ 将安全生产稳定性保障左移，在入网控制时介入，对入网控制、发布上线、故障预防、故障发现、故障定位，故障恢复、故障改进提供端到端工具支撑。



生产安全保障体系：一个目标，依托四大保障，聚焦研运流程中十二项核心工作，严格把控七个关口。

一个目标

做实安全生产，提升中国联通大IT系统稳定性

七个关口

设计关

验证关

上线关

变更关

监控关

应急关

优化关

十二项  
核心工作

架构设计

版本管理

上线交维

变更管理

监控管理

应急管理

故障管理

重保管理

混沌演练

隐患管理

容量管理

建维协同

四大保障

制度规范保障

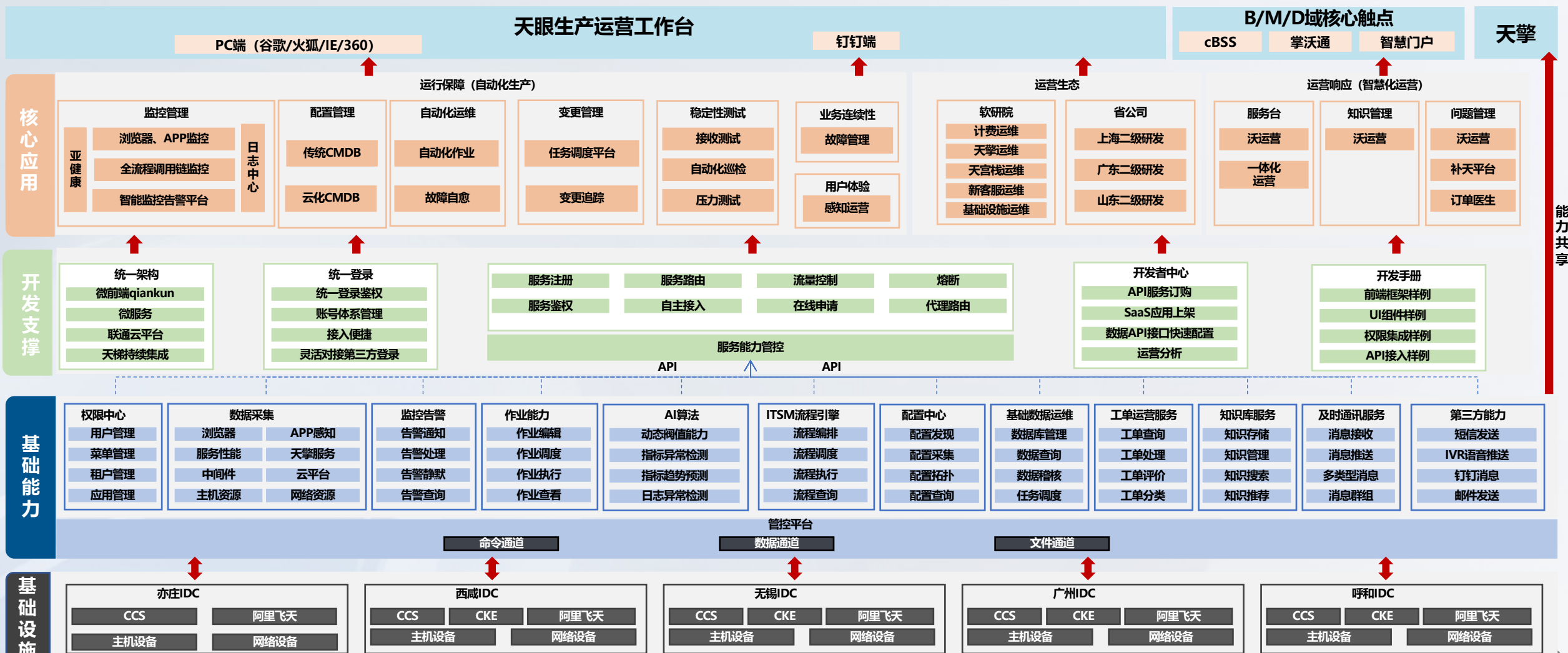
组织架构保障

平台工具保障

运营机制保障

# 数字化监控平台架构

基于云原生下的**生产运营支撑平台**，以全局运营视角解读IT运维，提供**端到端、全层级**的运维工具支撑，依托大数据与人工智能技术，助力企业数字化业务**高效、稳定运行**，从传统运维向**自动化生产、智慧化运营**转变。



03

稳定性保障的场景应用实践



- 稳定性保障核心场景要做到端到端的故障发现、故障定位、故障调度、故障处置、故障整改、故障预防。



## 及时发现

全层级实时监控，**1分钟**故障发现

## 智能定位

全链路深度追踪，**5分钟**故障根因定位

## 快速抢通

自动化应急预案，**15分钟**故障快速抢通

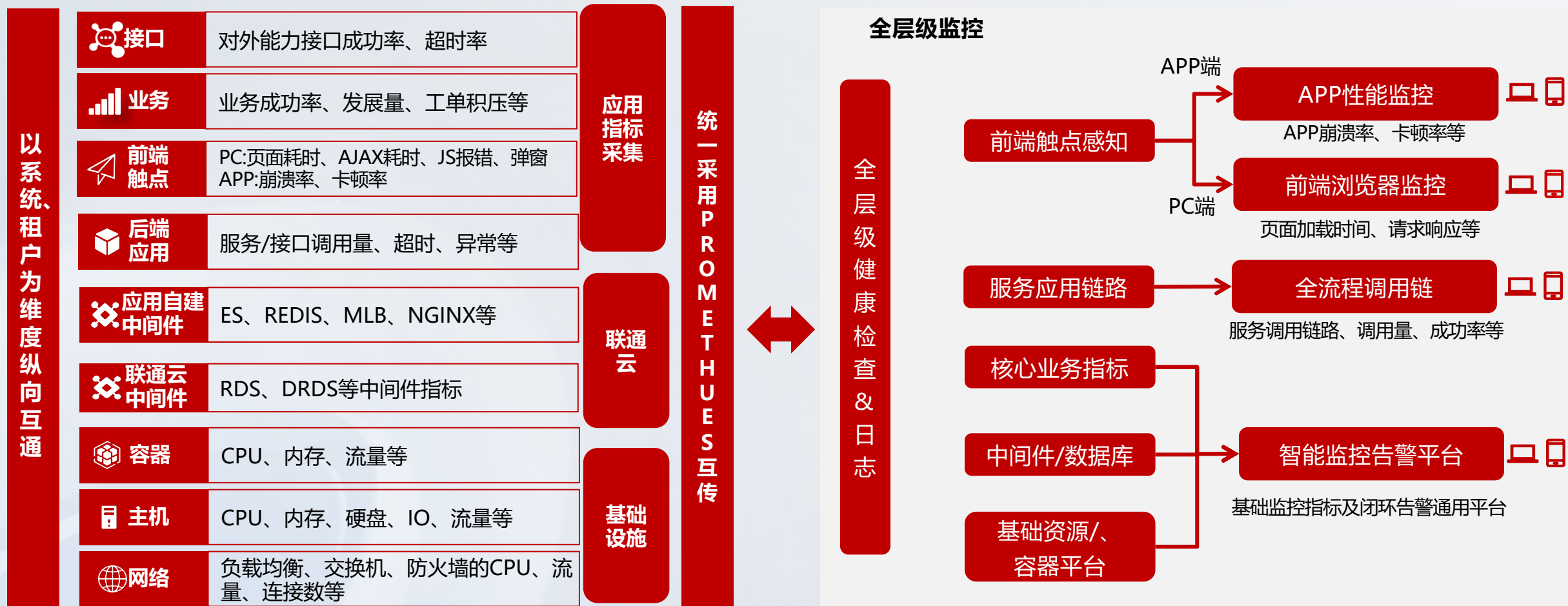
## 闭环治理

灵魂拷问，举一反三，**100%**故障闭环追踪

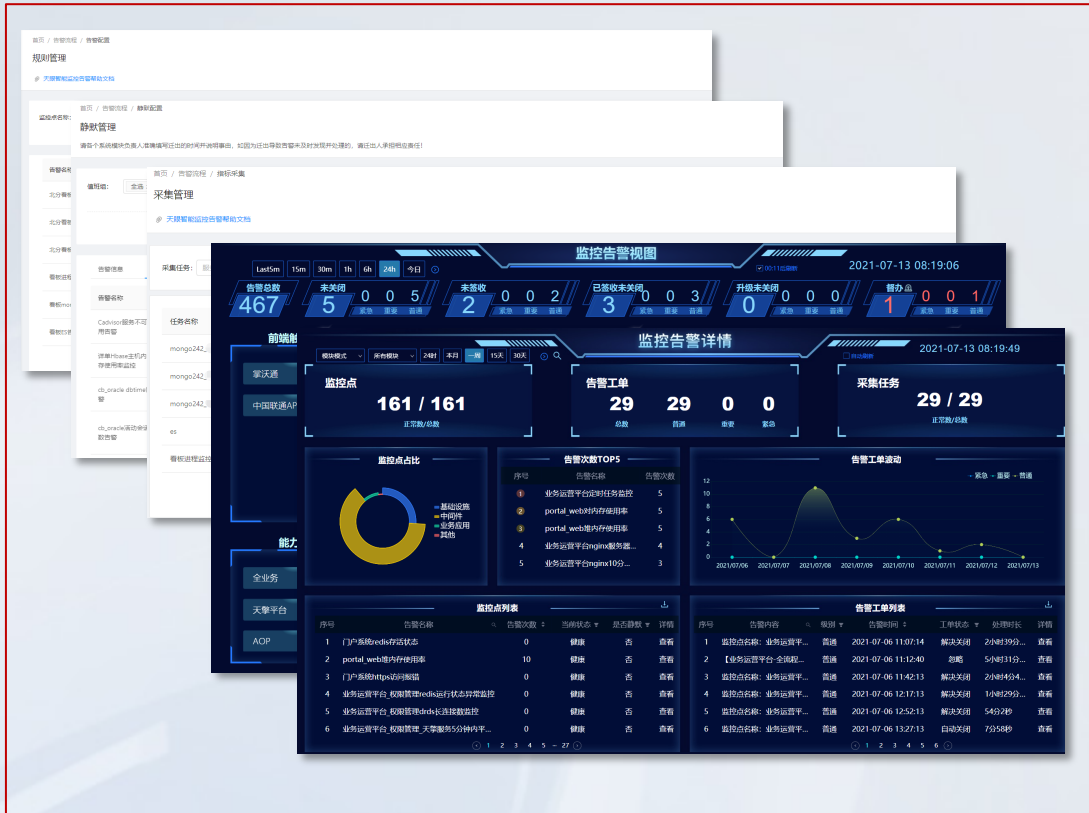
## 有效预防

应急演练、健康检查、智能自愈，**3重保障**

统一全层级监控标准，纵向互联互通，打破分散割裂格局，实现全层级、全链路、端到端的性能监控和链路追踪。



平台提供IaaS、PaaS、SaaS各层级监控能力，实现多层次运维数据互通，支持全流程可视化配置，多渠道告警通知，工单闭环管理，用户快速实现监控接入，为系统日常生产运行提供保障。

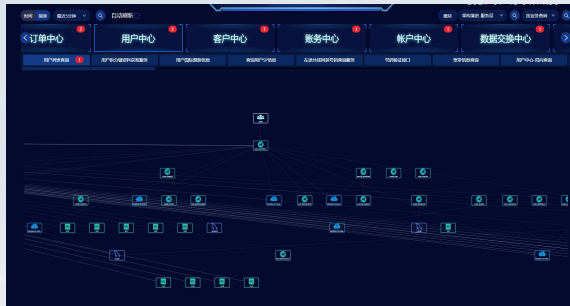


- 数据采集: 采集组件管理、私有数据仓库接入、租户自定义采集
- 监控配置: 告警规则、收敛条件、告警内容
- 静默管理: 多维静默管理 (全量、监控点、监控实例)
- 告警通知: 告警工单推送、电话催办
- 告警处理: 双终端工单处理、工单闭环管理
- 告警大屏: 系统监控告警全景图、告警工单处理进度

**制定全层级指标标准414项**

层级	类别	类型	问题	指标采集方式	指标类型 (中文)	指标类型 (英文)
14					服务器接收请求4XX数量	nginx_server_requests(code="4xx")
15					服务器接收请求5XX数量	nginx_server_requests(code="5xx")
16					后端转发耗时【单位毫秒】	nginx_upstream_requestMsec
17					后端转发4XX数量	nginx_upstream_requests(code="4xx")
18					后端转发5XX数量	nginx_upstream_requests(code="5xx")
19				天眼提供采集器 采集配置选择天眼 采集	当前正在处理的连接数	nginx_connections_current
20					抓取nginx时的错误数	nginx_exporter_scrape_failures_total
21					服务器接收请求总数	nginx_server_requests(code="total")
22					后端转发总数	nginx_upstream_requests(code="total")
23					连接数 (活跃)	nginx_server_connections(status="active")
24					连接数 (写)	nginx_server_connections(status="writing")
25					连接数 (读)	nginx_server_connections(status="reading")
26					连接数 (等待)	nginx_server_connections(status="waiting")
27	天官组件	SLB		天官阿里平台提供 指标 无需采集, 直接告 警配置	SLB实例状态	status_of_slb

通过探针非侵入式采集，实现调用链实时追踪、全层级故障根因定位。支持多租户、多系统接入、服务链路拓扑、多维根因定位分析、告警配置等功能。



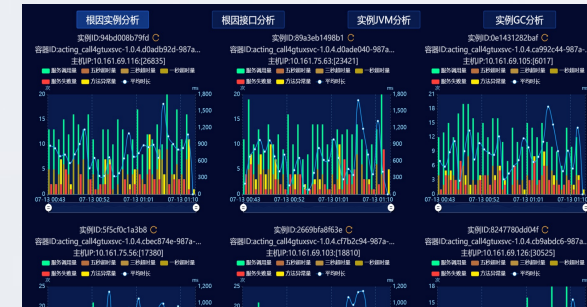
## 调用拓扑

全流程调用链拓扑自动生成，分租户管理



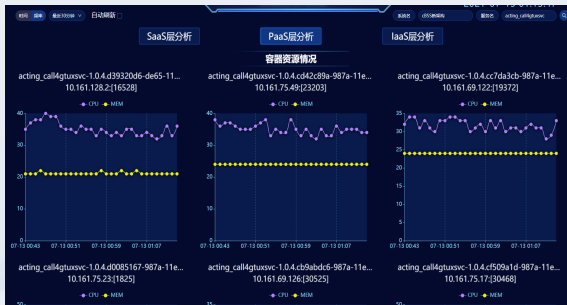
## 服务趋势/报错异常

服务调用关系、趋势图、报错分类（系统/业务）



## 实例/接口分析

调用链与云化CMDB做关联，关联到容器与主机



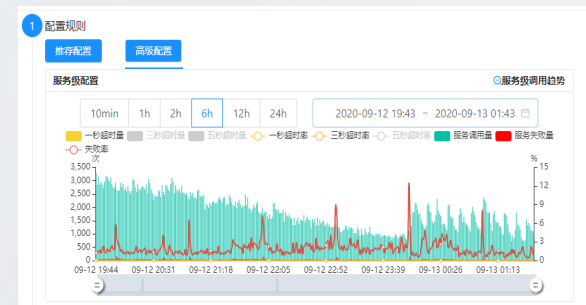
## SaaS/PaaS/IaaS

PaaS层组件、平台容器资源情况，IAAS层主机资源



## JVM/GC分析

服务实例JVM与GC情况分析

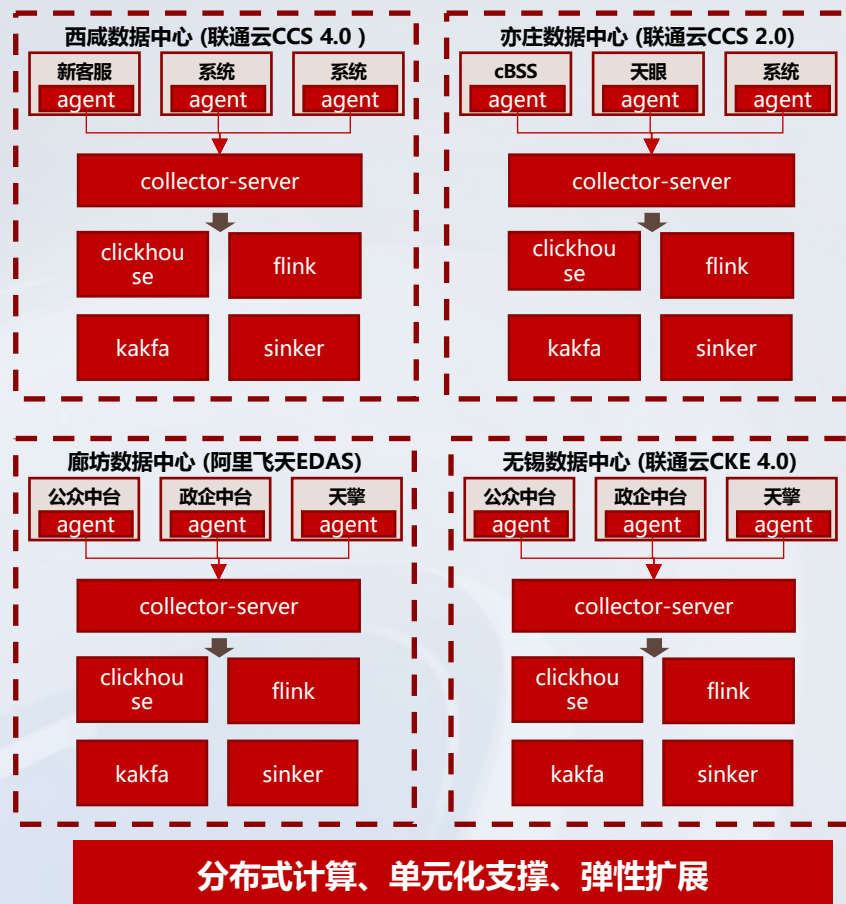


## 告警配置

调用量、超时、异常黄金指标多指标自由组合



支持跨系统、跨云平台（CKE/CCS/EDAS）、跨数据中心（亦庄、西咸、廊坊、无锡）链路拓扑，通过分数据中心汇总串联，完成跨系统调用实时追踪和方法清单级根因定位，日均处理近千亿数据。



创新点：跨数据中心链路组装



采用JS埋点的方式，采集用户访问过程的性能指标，获取浏览器端的真实用户行为与体验数据。包括页面加载、点击、弹窗、JS报错、ajax等用户全轨迹跟踪，通过大数据分析，应用于院内故障定位、安全分析、终端分析、感知分析、异常分析等场景。



系统总览



页面性能分析



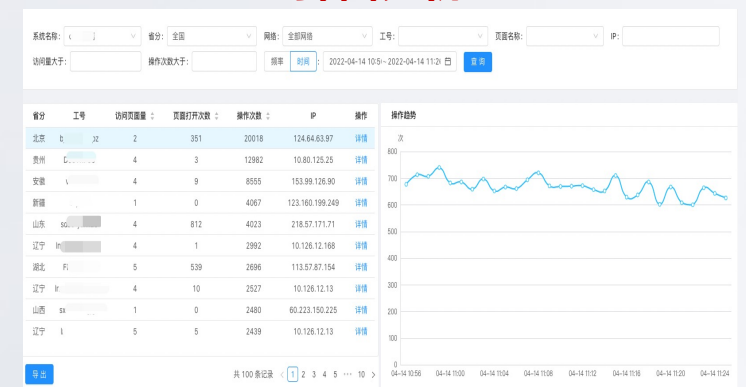
弹窗分析



AJAX分析



用户轨迹分析



工号稽核

通过采集指标、链路、报文日志，实现**三位一体**的可观测性，在系统纵向全层级方面实现触点层、服务层、组件层、平台层、主机层、网络层纵向贯通，结合云化CMDB关联定位，实现**全层级一键诊断**，端到端快速定位问题根因。

可观测

指标、链路、报文日志**三位一体**

智能诊断

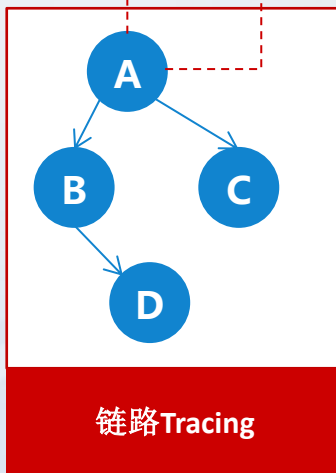
纵向贯通实现全层级**一键诊断**

指标  
Metrics

- ✓ 调用量
- ✓ 超时量
- ✓ 异常量
- ✓ 失败量
- ✓ ...

报文Logs

- ✓ 请求报文
- ✓ 响应报文
- ✓ 异常日志
- ✓ ...



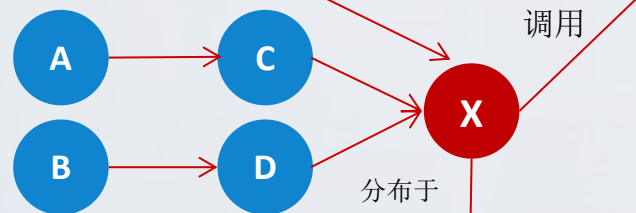
### 1. 发现业务影响

触点+业务监控评估影响范围。



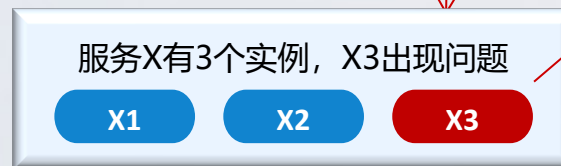
### 2. 定位根因服务

利用图数据库关系在海量告警服务中快速定位根因服务，如150个服务告警根因服务缩小到5个左右。



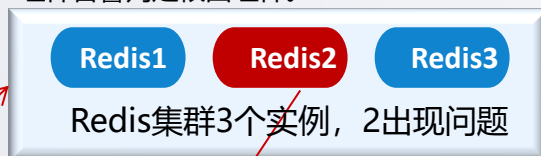
### 3. 定位根因实例

通过核密度估计算法和DBSCAN聚类算法判定根因实例。



### 4 定位根因组件

扫描根因服务调用的组件调用链指标、组件指标、组件告警判定根因组件。



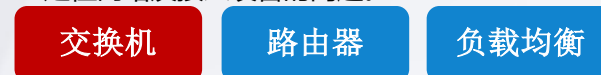
### 5. 定位根因主机

通过云化CMDB获取实例、组件与主机的关系，对主机的指标与告警进行扫描。



### 6. 定位根因网络

定位网络及接入设备的问题。





□ 依托全层级监控指标数据、全层级链路调用、云原生CMDB，建立故障传递模型，以服务层为故障起点进行纵向串联，配以规则+AI的能力实现全层级一键智能故障诊断。



服务实例异常：  
根因服务实例耗时突增  
实例GC引发故障



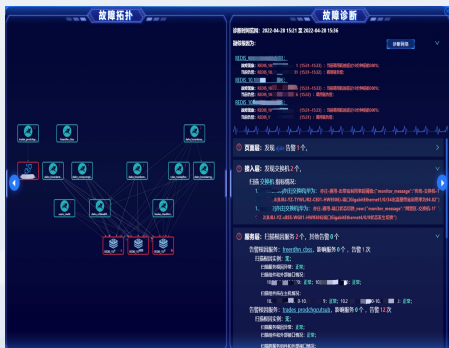
Oracle异常：  
Oracle会话数突增导致服  
务连接超时增多



ES异常：  
ES进程负载率突增导致上游  
服务连接超时



RDS异常：  
RDS慢sql突增导致节点状  
态异常



Redis异常：  
Redis耗时波动引起上游服  
务连接超时



快立方异常：  
根因服务下游调用快立方告  
警异常



主机宕机异常：  
lb所在主机宕机导致lb实  
例销毁重启服务波动



网络异常：  
网络带宽使用率指标打满  
引起访问受限

## ■ 全层级指标数据

- ✓ 分布式链路拓扑数据
- ✓ 全层级核心监控指标

## ■ 云原生CMDB

- ✓ 服务、组件、主机、网络关系拓扑

## ■ 以服务为起点纵向关联

- ✓ 云原生下以服务告警触发进行上下游关联

## ■ 智能根因定位

- ✓ 服务异常实例波动
- ✓ 平台组件指标异常
- ✓ 主机异常宕机夯死
- ✓ 网络设备带宽打满



- 将“监”、“管”、“控”工具能力融合，告警信息结合AI判定算法，触发自动化作业能力，实现故障自愈流程，有效缩短故障处理、恢复时间。



## 16:20:45 应用告警



告警查询

工单编号

编号: 085649735781

告警信息

租户	模块	维保组	级别	重要	检测时间
					2023-09-07 16:20:45

详细内容: [租户] - 全流程调用链告警,近1分钟【发现时间:2023-09-07 16:18:59, 根因服务: [租户], 系统异常量:84, 系统异常量环比昨日:8300.0%, 系统异常量环比上一分钟:82.61%[此项超出阈值, 告警阈值:系统异常量>15且系统异常量环比昨日>100%且系统异常量环比上一分钟>0%], 疑似根因异常: [租户], 微服务研发负责人: [租户]

处理历史

- 2023-09-07 16:20:47 系统告警发起外呼: [租户]
- 2023-09-07 16:22:02 [租户] 已签收
- 2023-09-07 16:22:11 [租户] 解决关闭

## 16:20:47 推送实例查杀、重启工单



自愈历史详情

套餐名称: [租户]单实例异常自动kill自愈套 作业名称: [租户]单实例异常自动kill 策略名称: [租户]单实例异常自动kill策略

发生时间: 2023-09-07 16:20:46 匹配成功时间: 2023-09-07 16:20:47 作业执行成功时间: 2023-09-07 16:21:34

审批成功时间: 2023-09-07 16:21:29 自愈耗时: 3530 审批人: [租户]

审批工单: [租户] 作业参数: [租户] 结果: 作业执行成功

事件详情

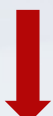
告警名称: [租户] - 全流程调用链告警,近1分钟 告警源: 调用链 告警级别: 主要告警

告警内容: [租户] - 全流程调用链告警,近1分钟

系统: [租户] 模块: [租户] 告警时间: 20230907162045839

异常实例: [租户]

告警信息: [租户] - 全流程调用链告警,近1分钟【发现时间:2023-09-07 16:18:59, 根因服务: [租户], fo, 系统异常量:84, 系统异常量环比昨日:8300.0%, 系统异常量环比上一分钟:82.61%[此项超出阈值, 告警阈值:系统异常量>15且系统异常量环比昨日>100%且系统异常量环比上一分钟>0%], 疑似根因异常: [租户], 微服务研发负责人: [租户]



## 16:20:45 自动触发诊断



故障诊断

系统名称: [租户]

诊断时间范围: 2023-09-07 16:18 至 2023-09-07 16:21

正在进行诊断...

服务告警趋势: [租户]

页面层: 发现 页面告警 1 个, 扫描 页面情况, 告警次数 1 个

全部页面内容: 告警次数 1 个, 告警信息: 起始告警时间: 2023-09-07 16:20:00 最近告警时间: 2023-09-07 16:20:00 实例名称: [租户]

接入层: 检查Marathon-LB\_Kong, 未发现问题

服务器层: 扫描根因服务 1 个, 其他告警 0 个

告警根因服务: [租户] 影响服务 0 个, 告警 3 次

扫描根因实例: [租户]

扫描实例状态: 正常

扫描实例所在主机: [租户]

扫描主机问题: 正常

组件层: 未发现问题



## 16:21:29 运维人员确认操作

## 16:21:34 应用恢复



根因分析视图

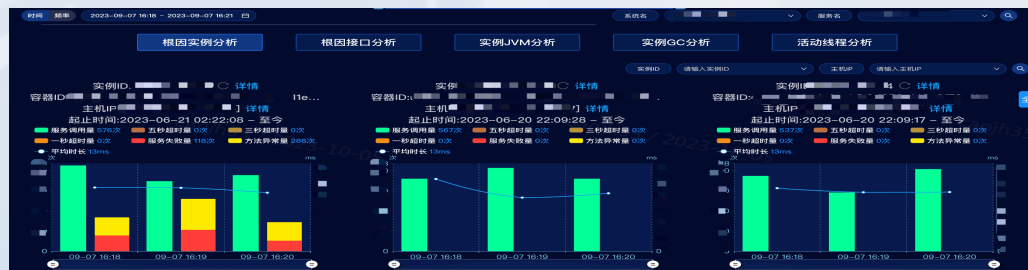
时间: 2023-09-07 16:21 - 2023-09-07 16:26 日

SaaS层分析 PaaS层分析 IaaS层分析

服务调用关系

调用量: [租户] 一秒钟时量: 0 三秒钟时量: 0 五秒钟时量: 0 服务失败数量: 0 系统异常量: 0

被调用量: [租户] 一秒钟时量: 0 三秒钟时量: 0 五秒钟时量: 0 服务失败数量: 10 系统异常量: 24



根因实例分析

容器ID: [租户] 实例ID: [租户] 详情

主机IP: [租户]

起止时间: 2023-06-21 02:22:08 - 至今

服务健康: 100% 二秒超时量: 0% 三秒超时量: 0% 一秒钟时量: 0% 服务失败数量: 0% 方法异常量: 0%

平均时长: 13ms

容器ID: [租户] 实例ID: [租户] 详情

主机IP: [租户]

起止时间: 2023-06-20 22:00:28 - 至今

服务健康: 100% 二秒超时量: 0% 三秒超时量: 0% 一秒钟时量: 0% 服务失败数量: 0% 方法异常量: 0%

平均时长: 13ms

容器ID: [租户] 实例ID: [租户] 详情

主机IP: [租户]

起止时间: 2023-06-20 22:09:17 - 至今

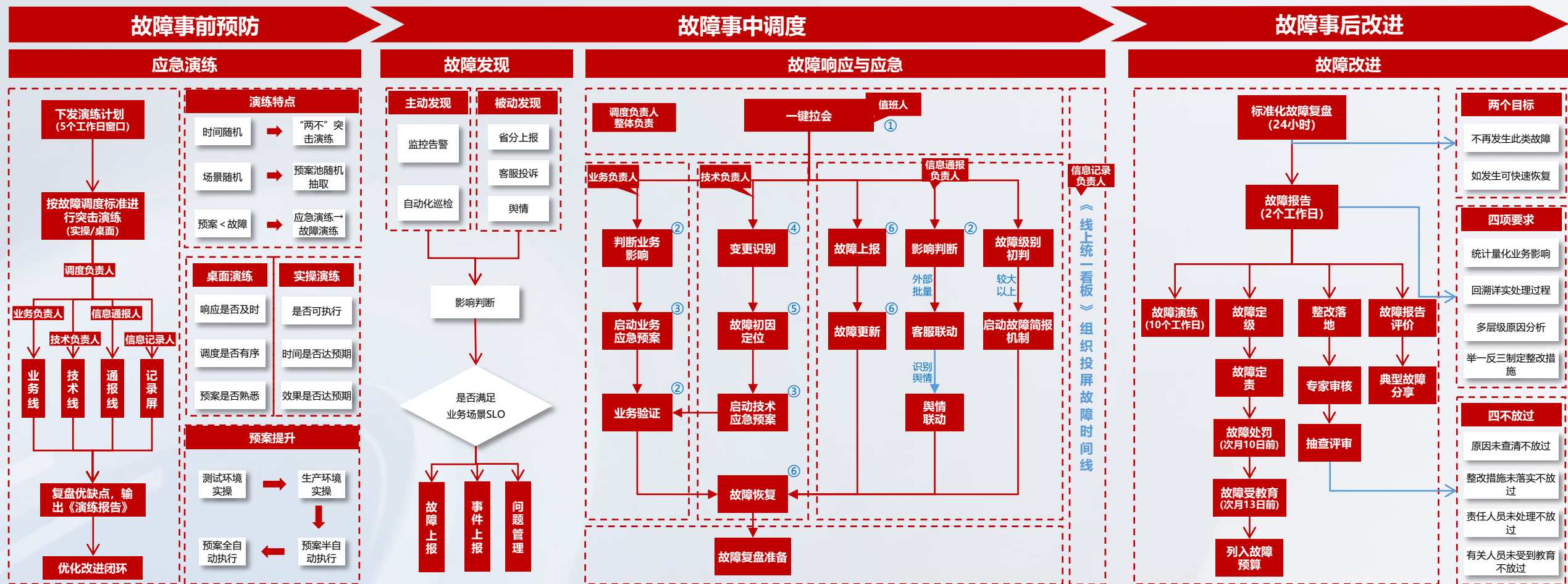
服务健康: 100% 二秒超时量: 0% 三秒超时量: 0% 一秒钟时量: 0% 服务失败数量: 0% 方法异常量: 0%

平均时长: 13ms



节省4分钟 → 从收到告警到恢复仅用47s

故障事前、事中、事后全流程线上闭环管理，提升故障管理质量和效率，降低故障时长及次数，提升业务连续可用率。

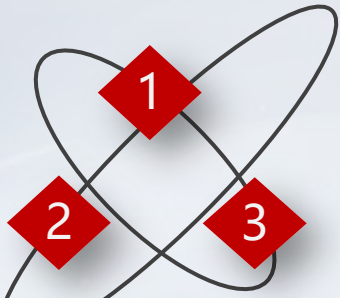


结合监控指标与容量指标，定期开展容量隐患评估，通过核心业务链路的全链路压测，分析链路性能瓶颈，建立健康度算法模型，识别与治理系统潜在风险隐患，保障系统健康稳定。

## 容量隐患分析

### 容量标准制定

- 业务、服务、组件、基础资源容量水位模型



### 目标容量评估

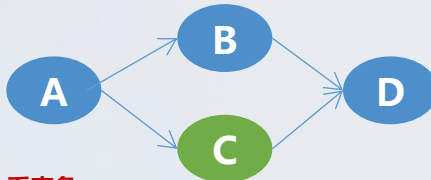
- 全链路压测 -> 容量标准达标、链路性能瓶颈评估...
- 日常流量方法级分析 -> 抖动、不达标率...
- 指标实时监控 -> 容量风险监控...
- .....

### 容量问题优化

- 对照保障目标，形成容量优化提升项
- 制定容量优化方案计划
- 容量再评估直至符合预期容量标准要求
- .....

## 链路性能瓶颈分析

### 定位性能瓶颈节点初步定位



#### 看表象

- 链路节点RT增长->初步定位瓶颈节点
- trace明细分析 -> Gap等待时间长、自耗时高、慢SQL...
- 链路调用量 -> 重复调用问题
- .....

#### 解释表象

- 线程池、连接池是否打满
- 慢SQL分析
- 内部方法自耗时高原因
- 重复调用是否可优化
- .....

#### 性能治理

- 压测结论、问题、论证、优化方案
- 与研发侧确认问题、推动治理
- 流量回放与复测等
- .....

### 深度性能问题分析

### 整体报告生成推动治理

## 系统健康检查

### 健康检测引擎

#### 页面层检测

- 页面弹窗数
- 页面JS错误
- 页面平均响应时长
- .....

#### 服务层检测

- 服务超时率
- 服务异常率
- 服务调用量
- 服务平均响应时长
- .....

#### 组件层检测

- ES健康节点/堆内存使用率/...
- REDIS内存使用率/内存碎片比率
- KAFKA消息积压/topic副本
- .....

#### 资源层检测

- 内存使用率
- cpu使用率
- 磁盘使用率
- .....

### 评分与趋势

#### 高风险指标

#### 中风险指标

#### 低风险指标

### 运营闭环管理

#### 实时监控体检

#### 隐患报告定时推送

#### 风险问题闭环整改

#### 性能对比



自动获取全层级**核心黄金指标**，通过AI算法分析，优化健康度算法模型，进行**全层级隐患分析**，实现系统**健康状态档案化管理**，分析与治理**潜在风险隐患**，保障**核心业务连续性**。



### 周期性观测

故障预防统计以日、周、月维度统计问题项情况，观测系统阶段性运行情况



### 实时健康体检

系统实时体检实时计算全层级指标，根据阈值判断指标异常及风险程度



### 性能对比

系统性能对比页面可选取发版前后时间进行各指标性能对比，观测系统性能变化趋势

### 3.三天亚健康实体统计

周期性实例，请重点关注。

实体	指标名称	20220401		20220331		20220330		三天亚健康次数	三天亚健康率
		亚健康次数	亚健康率	亚健康次数	亚健康率	亚健康次数	亚健康率		
prod-gdhh/node-6-exp1	10分钟qct时长增量	89	100.0%	94	100.0%	70	88.0%	253	96.0%
prod-gdhh/node-7	10分钟qct时长增量	83	93.0%	93	98.0%	68	88.0%	244	93.0%
prod-gdhh/node-5	10分钟qct时长增量	88	98.0%	80	86.0%	69	88.0%	237	91.0%
prod-gdhh/node-6	10分钟qct时长增量	89	100.0%	68	73.0%	77	98.0%	234	90.0%
prod-gdhh/node-7-exp1	10分钟qct时长增量	87	97.0%	67	71.0%	73	92.0%	227	86.0%

### 体检报告

### 健康检查指标分析报告-组件层

体检时间: 2022-03-28 09:00:00至2022-04-01 17:00:00

目录

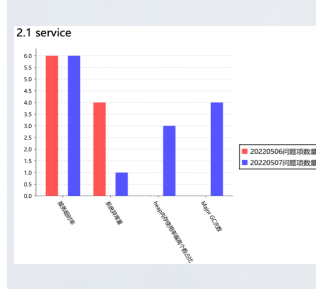
- 基础信息
- 总体结论
- 三天亚健康统计
- 亚健康指标详情

1.基础信息

2.总体结论

2.1 总体情况

问题名称	问题个数	问题个数占比
ms	5	41.67%
api	0	0.0%
entity	0	0.0%
mq	3	25.0%
spark	0	0.0%
redisearch	3	25.0%
redis	1	8.33%



实体问题列表

优化建议: 建议查看性能优化

实体	问题名称	亚健康次数	异常总次数	问题	平均分	最低分	最高分
misc_kurhmc_hydrx_db		14	41	服务Major GC耗时超过1000ms实例数...	82.93	50	100
freedjs_paperless		15	49	服务Major GC耗时超过1000ms实例数...	84.69	50	100
freedrn_rtd		16	57	服务Major GC耗时超过1000ms实例数...	85.61	40	100
freedrb_param		2	7	服务Major GC耗时超过1000ms实例数...	85.71	50	100
freedrb_sysnum		15	56	服务Major GC耗时超过1000ms实例数...	86.61	50	100
freedrb_apivcrum		11	42	服务Major GC耗时超过1000ms实例数...	86.9	50	100
freedrdl_smpatform		5	23	服务Major GC耗时超过1000ms实例数...	89.13	50	100
freedrb_param		6	32	服务Major GC耗时超过1000ms实例数...	90.31	40	100
cd1_cms_pty_basereport		9	57	服务Major GC耗时超过1000ms实例数...	91.93	40	100
numport_status_update		1	7	服务Major GC耗时超过1000ms实例数...	92.86	50	100

### 黄金核心指标选择

根据故障知识库与专家建议，选取页面、服务、组件、资源层核心黄金指标

系统体检与性能对比报告，找出系统异常指标标注指标含义、可能引起故障、整改举措，助力系统整合，夯实稳定性

# Q&A



<https://sre-elite.com>

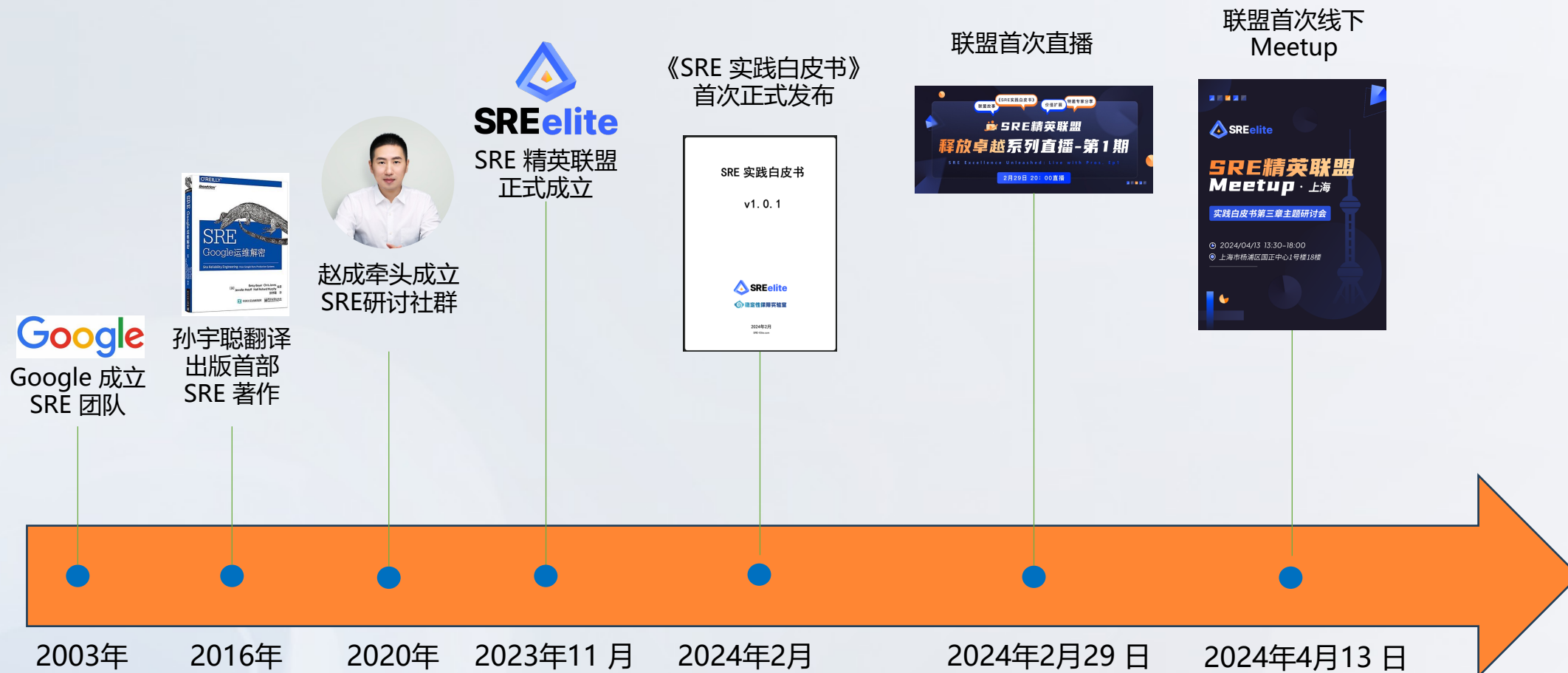
# 附录

供Meetup主持人和分享嘉宾参考



<https://sre-elite.com>

# “SRE精英联盟”概述





SRE 实践白皮书

v1.0.1



2024年2月  
SRE-Elite.com



经历数年，20 多位一线专家协作编写。



扫码下载 v1.0.1。版本持续更新迭代中。



在官网 <https://sre-elite.com/notice/> 下载最新版。



公众号



视频号



B 站



YouTube