全球化游戏故障管理实践

2024年6月

https://sre-elite.com

个人介绍

姓名 乾海平

腾讯自研游戏SRE负责人

工作经历：

2010-2024年，腾讯科技

2004-2010年，深圳电信

个人简介：

2010年加入腾讯，负责多款大型头部自研和代理业务的运营规划工作，先后在容器平台、CI/CD，混合云管理，SRE等项目上，成功推动专项解决方案落地。擅长游戏业务全生命周期的服务规划和解决方案实施，尤其是对海外游戏全球化技术运营有深刻理解。目前正专注于腾讯游戏的SRE体系建设。

目录

CONTENT

SREelite

01

探索之旅：全球化游戏业务面临故障管理的问题和挑战

SREelite

全球化游戏业务与国内游戏业务的主要区别

主要区别：

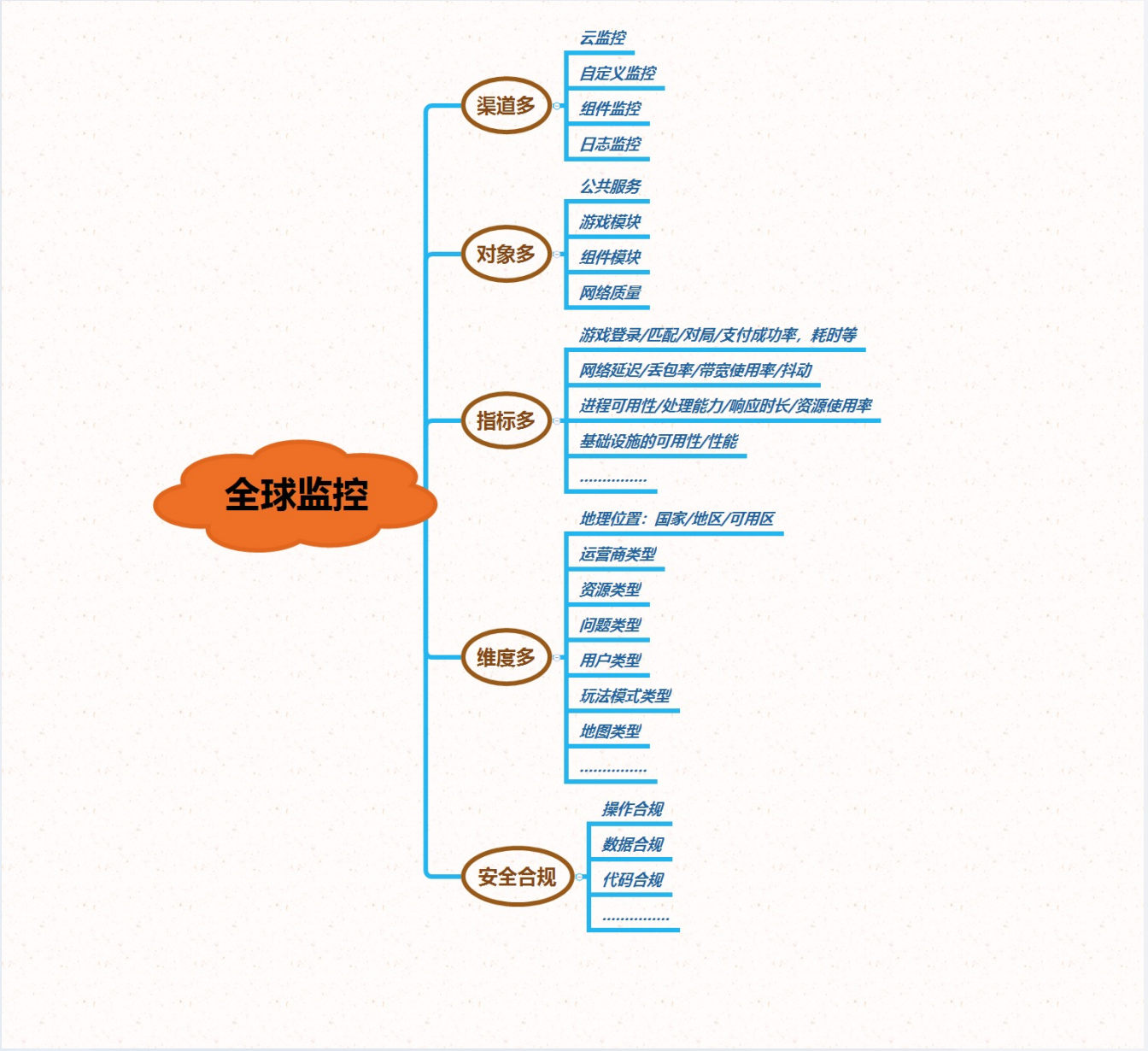
范围	网络	监控指标	文化差异	云资源	法律政策	时区差异
全球化业务	网络环境复杂	指标和维度更多	文化多样	混合云部署	法规多元	多时区
国内业务	网络环境相对稳定	指标和维度相对单一	文化差异不大	单云部署	法规统一	单时区

典型的区域式全球同服架构特点

用户分布在全球各地
战斗集群就近部署
公共模块集中部署

全球化游戏监控面临的挑战

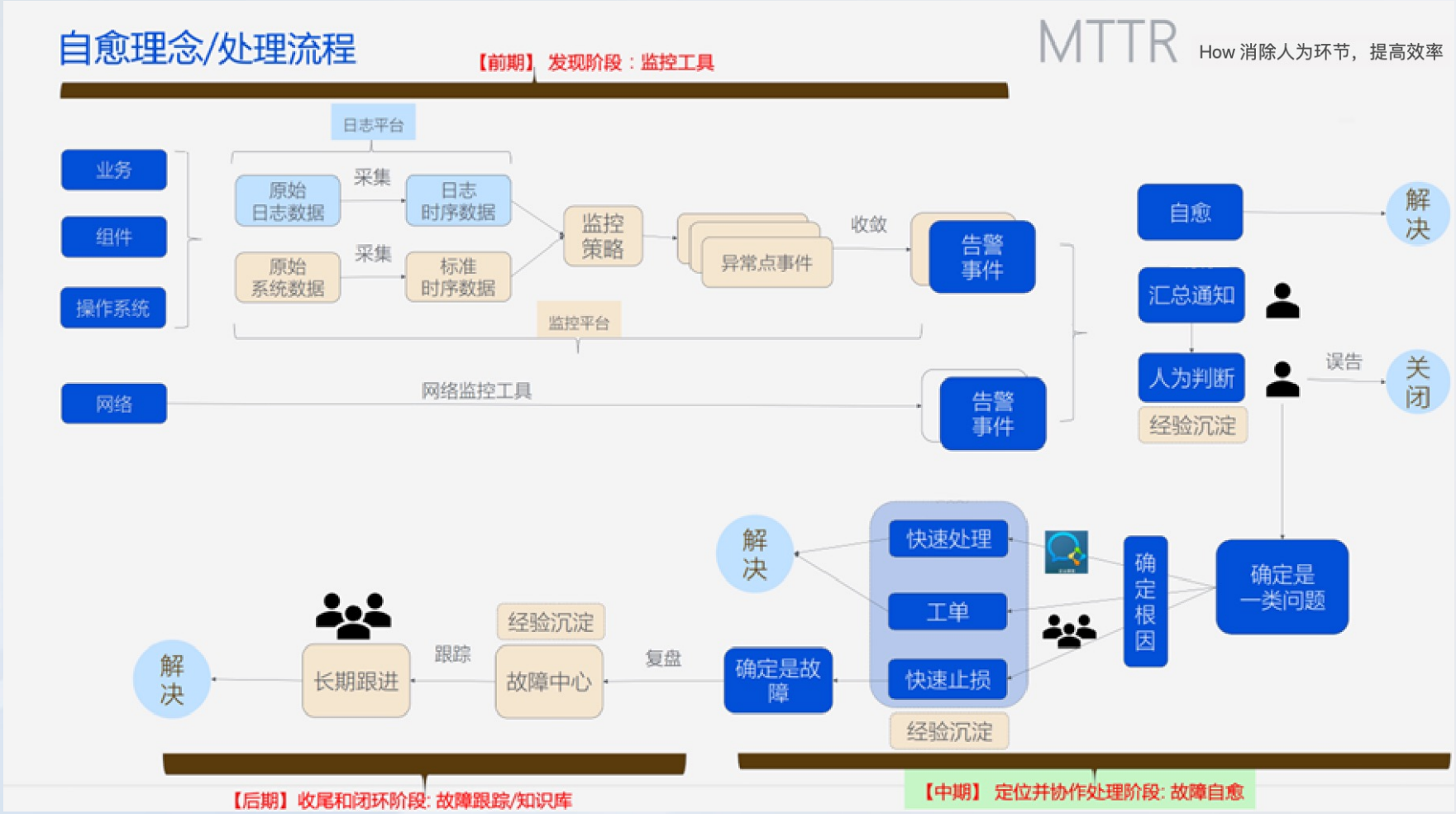
全球化游戏监控要面临的挑战：
复杂的系统多样性
数据渠道碎片化
多维度指标
严格的安全合规要求



SRE故障管理的全链路流程：从发现到改进

SREelite

- 提高系统的可靠性和稳定性
- 提高服务质量和用户满意度
- 提高组织的效率和生产力
- 提供学习和改进的机会



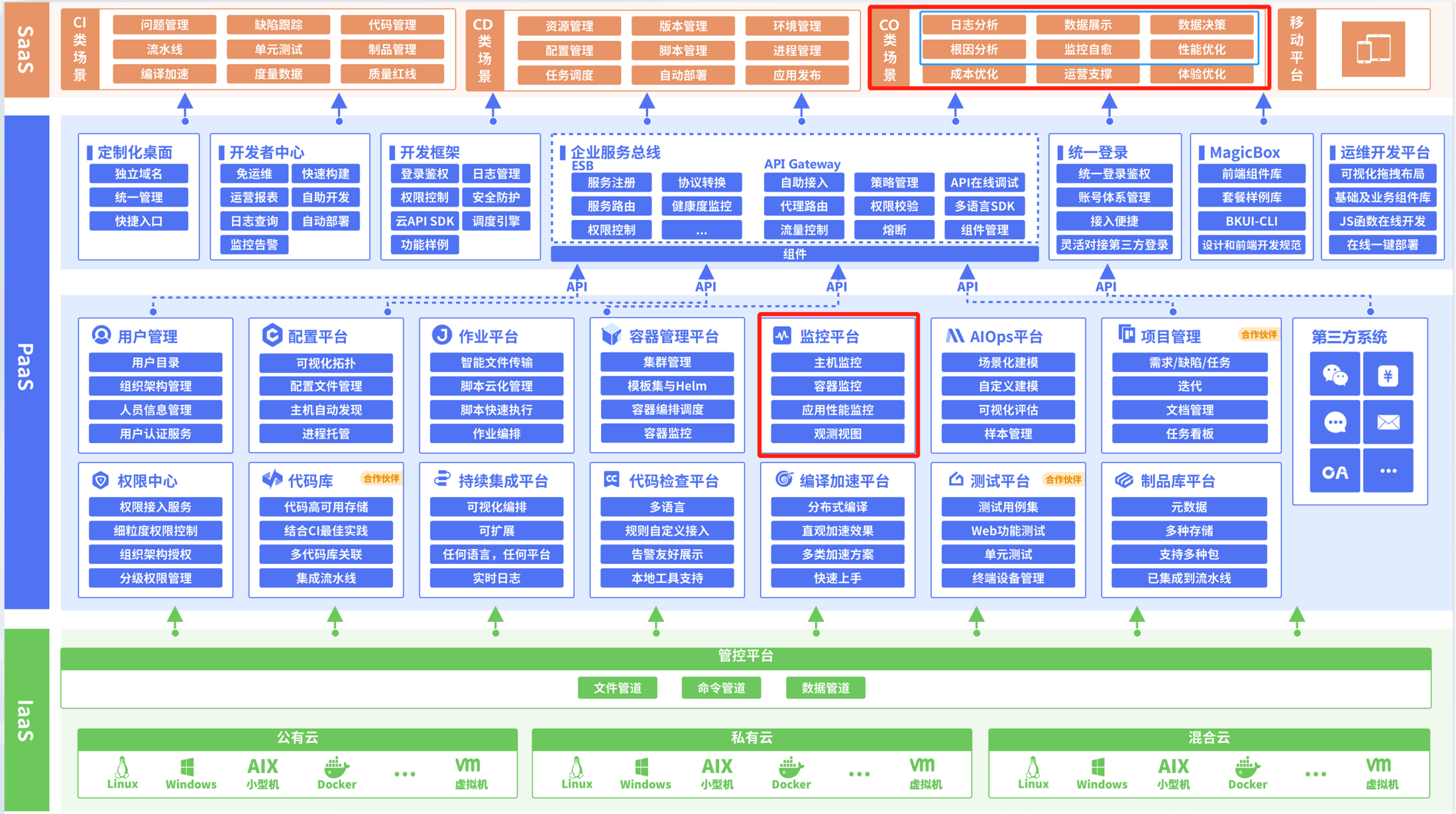
02

破解密码：如何有效发现并定位故障

SREelite

依托蓝鲸生态，打造全球化的故障治理体系

SREelite



蓝鲸监控全球支撑模式

SREelite

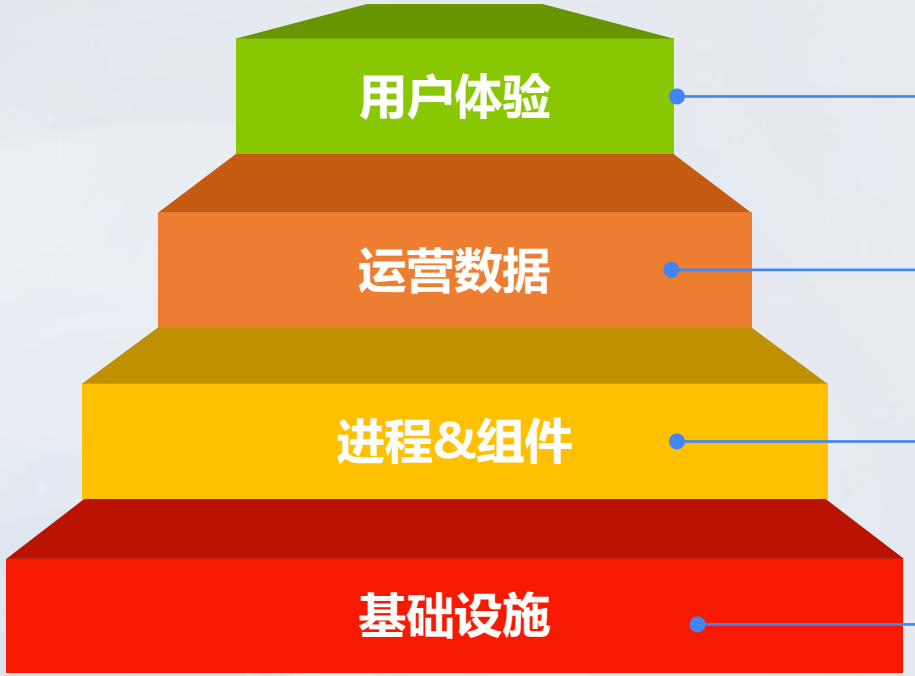
混合云

全球化的监控体系
全球覆盖
混合云部署

Step1 ： 基于蓝鲸监控体系下，业务指标如何管理-指标分层

SREelite

指标分层



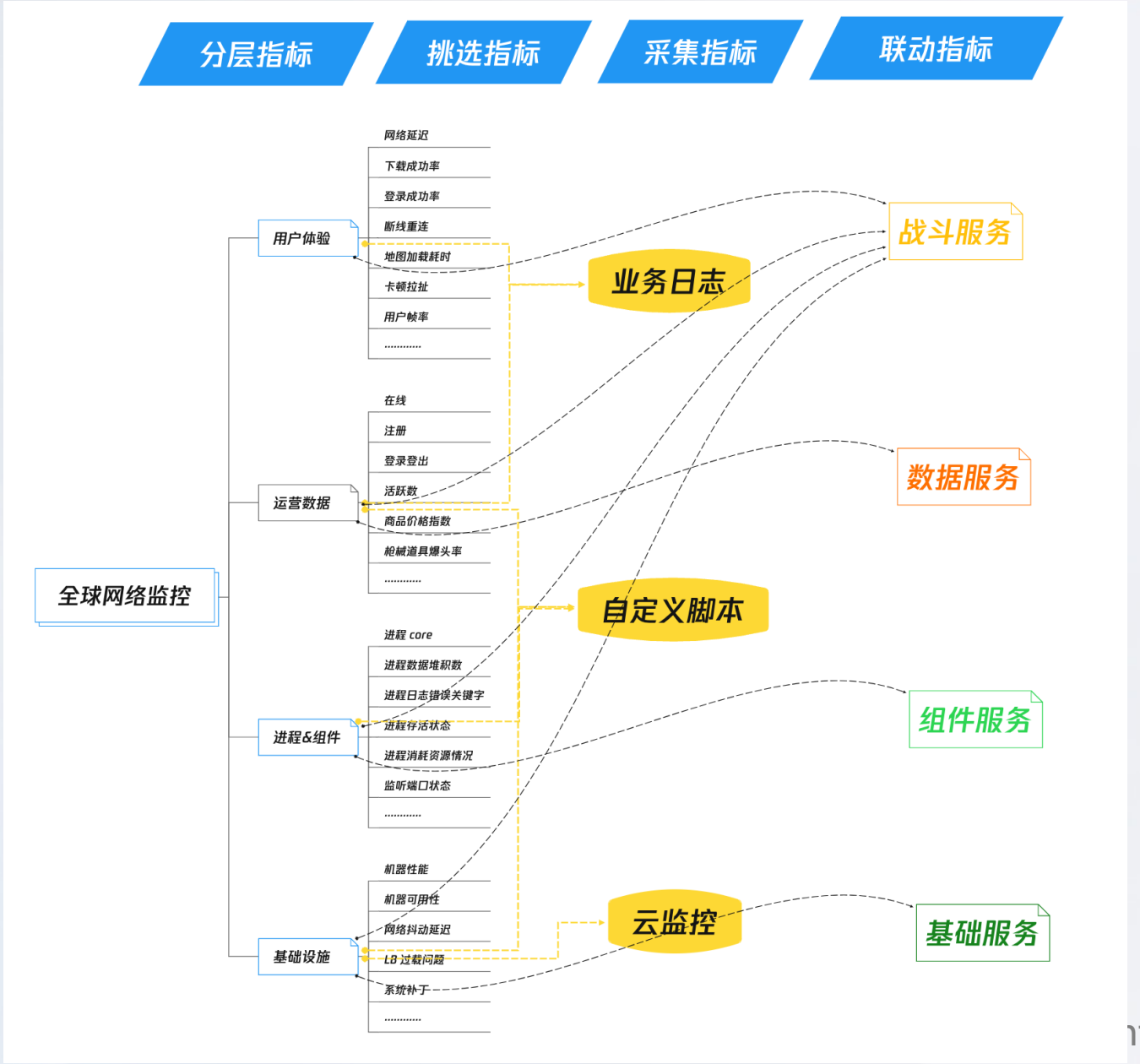
和用户强相关的指标
如：对局网络延迟，登录和匹配成功率等

和运营强相关的数据
包括：在线，注册，登录登出

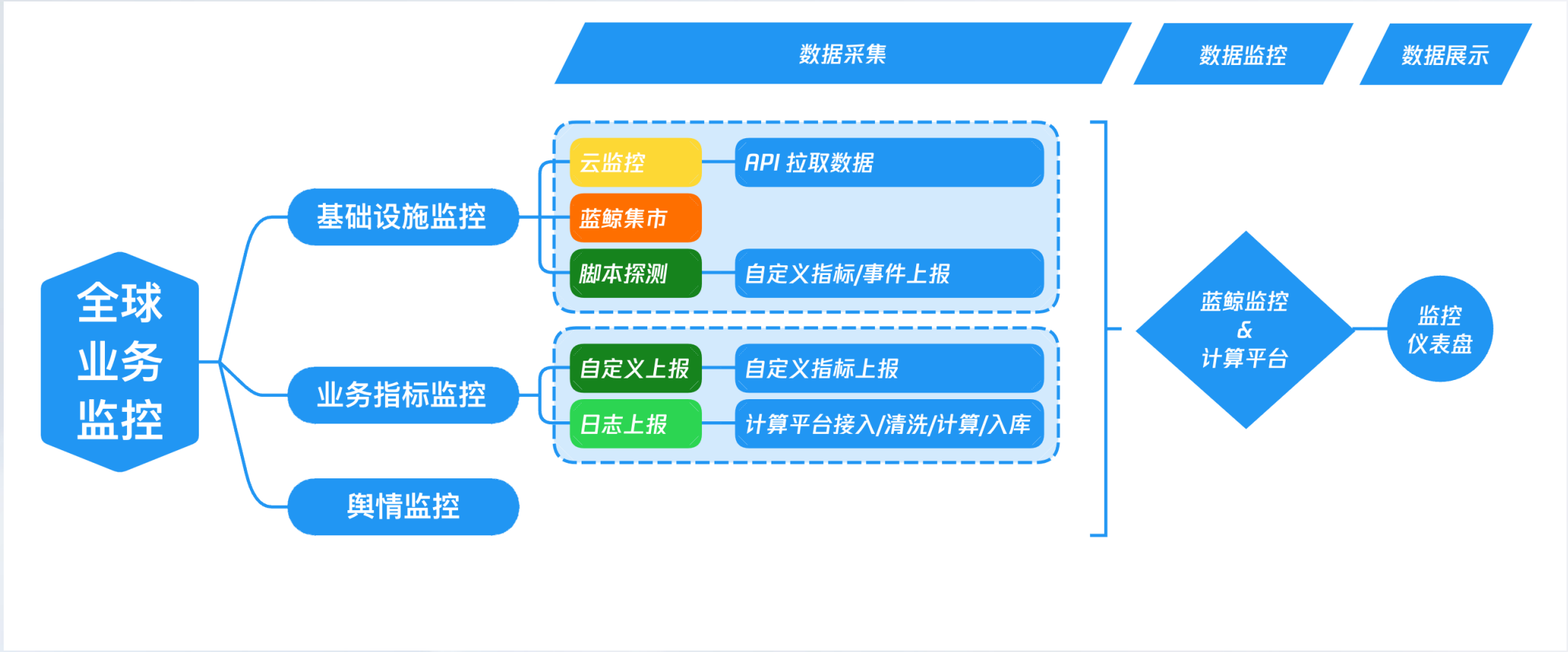
服务端程序本身，以及程序所依赖的组件
包括：游戏业务进程，大区列表，支付，下载类等关键组件

云基础设施
网络类，机器类，存储类等

Step2：建立监控指标体系



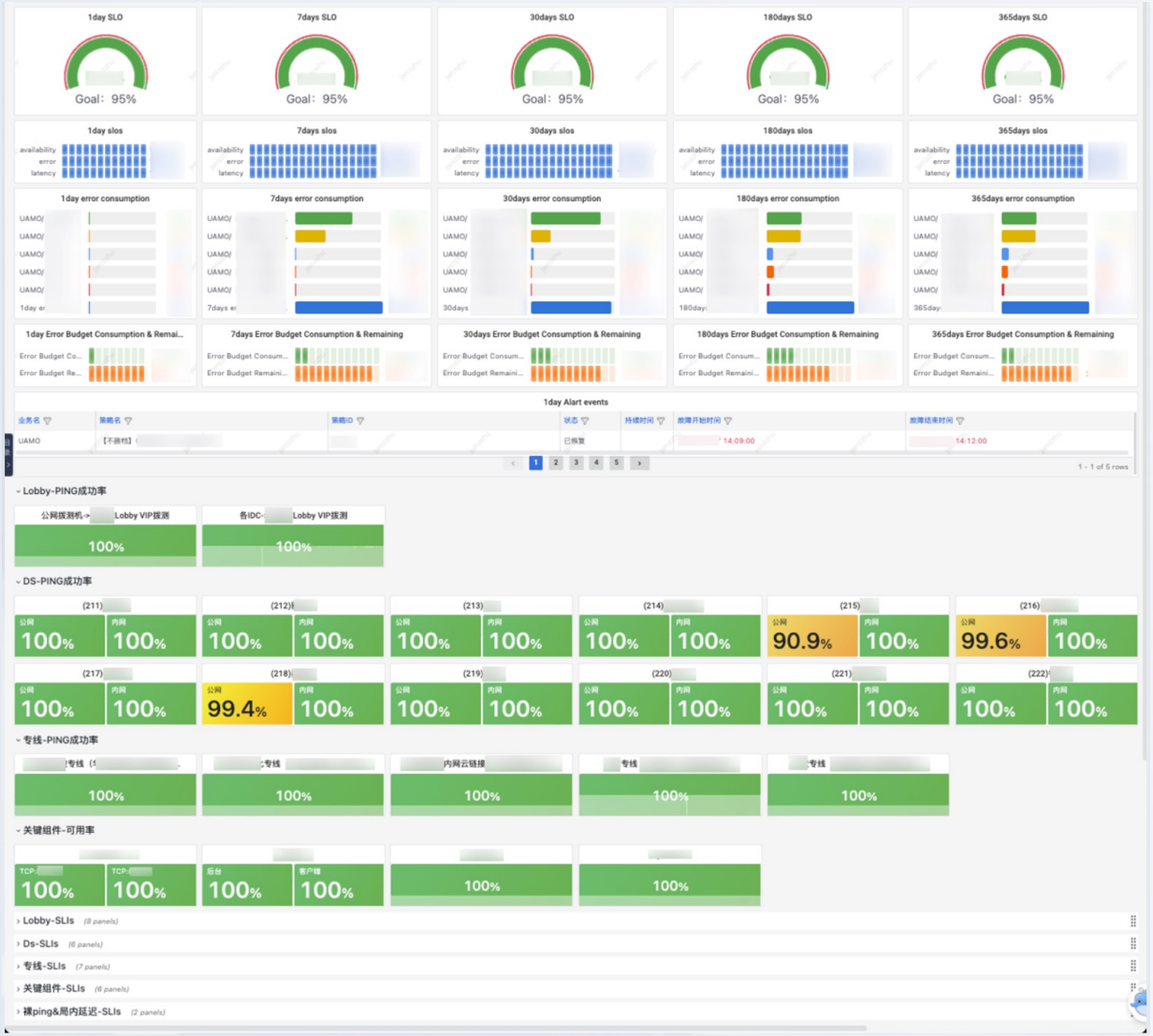
Step3：数据采集，统一管理



Step4：建立网络质量的可观测大屏，快速发现和定位问题

SREelite

总体到局部的分层展示
基础设施到游戏应用可用性展示
展示**全局**网络质量SLO
游戏模块的ping成功率
专线ping成功率
关键组件的ping成功率
裸ping局内延迟



蓝鲸监控仪表盘显示的SLO

Step4-1：基础设施可观测大屏

SREelite

蓝鲸拨测-轻松了解关键组件网络质量

下载服务可用性
登录，大区列表服务可用性
支付服务可用性
排行版服务可用性

Step4-2：业务对局体验可观测

SREelite

对局网络质量分析

通过热力图，能够告诉我们哪些IDC网络是最佳的，哪些又是最差的，为我们的云节点选择提供决策

IDC的热力图

玩家分配IDC热力图

案例1：维度下钻和数据关联，定位菲律宾网络问题

SREelite

e.g. 晚上黄金时分，菲律宾玩家网络延迟为何突然升高?

1. **告警触发**：菲律宾最优 IDC 对局延迟升高 -> **触发监控告警**

2. **维度下钻**：检查**分IDC对局延迟**时序图 -> **对局分配重点IDC中**，延迟均升高且趋势一致
-> **基本排除 IDC 问题**

3. **关联检查**：检查对局样本数量 -> 样本数量无异常
检查网络质量拨测数据 -> 无异常
检查菲律宾在线数量 -> 在线无影响
检查裸 ping 延迟 -> **有同样上升趋势**

-> **基本排除业务网络链路问题**，怀疑是公共网络故障

4. **换一个维度下钻**：检查分运营商对局延迟时序图 -> **有两个TOP运营商延迟上升趋势和总体上升趋势一致**

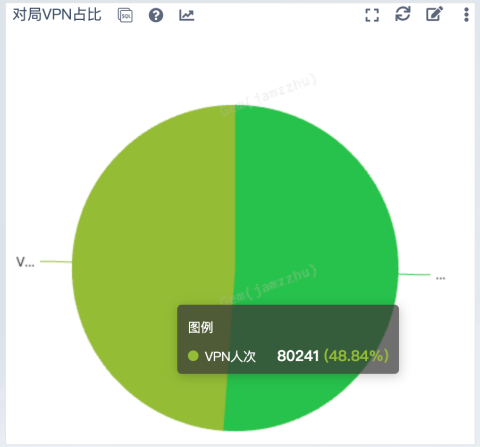
-> **初步判断为运营商问题**

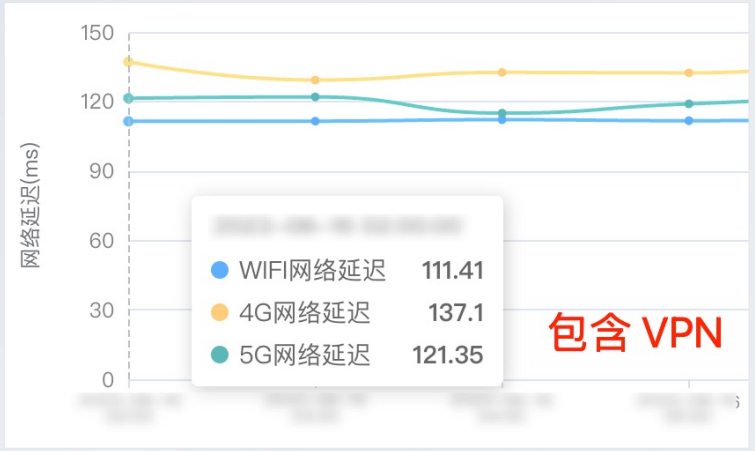
5. **公共网络问题求证**：找云厂商核实 -> **确认为运营商光缆故障导致**

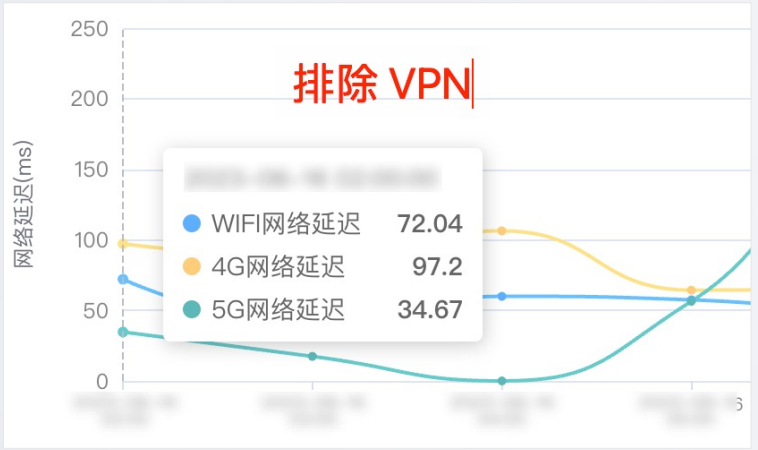
案例2：网络可观测，数据层层剥离，破解网络疑难杂症，提供运营决策

SREelite

- **新加坡玩家明明网速流畅，为何大盘延迟居高不下？**



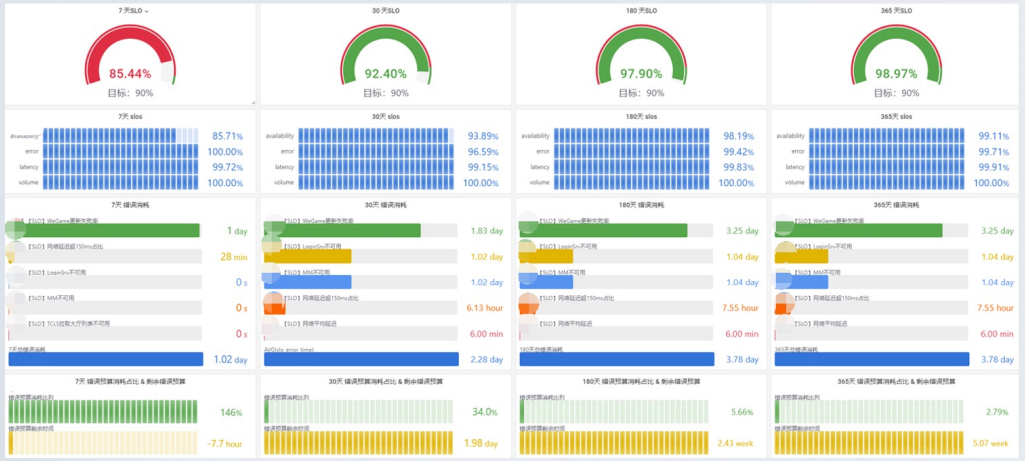




- **伊拉克玩家为何不就近匹配，反而绕路去了欧洲节点？**

案例3：业务关键路径质量可观测，快速发现登录问题

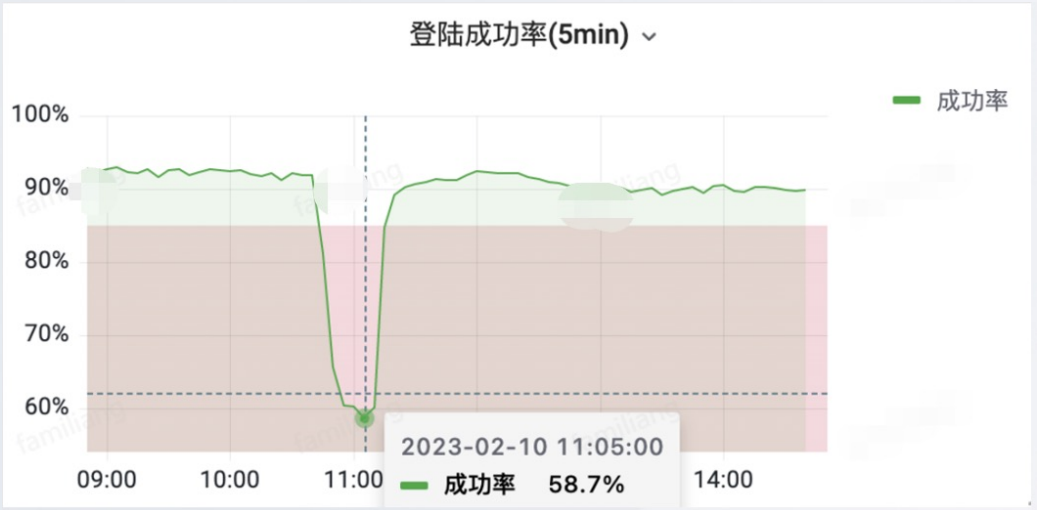
SREelite

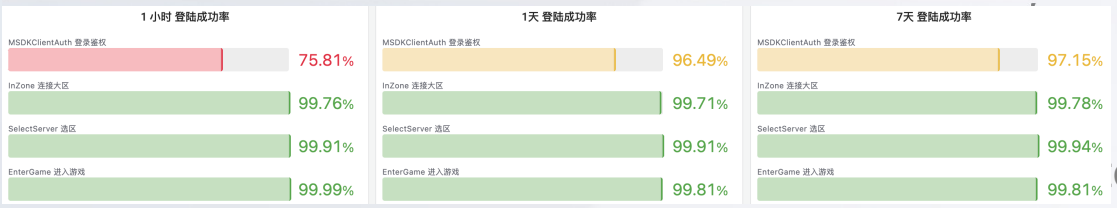


关键节点

Time	1、msdk鉴权成功率	3、Lobby进程存活性	4、Lobby主机CPU使用率	4、Lobby主机MEM使用率
2023-02-10 11:30:00	98.68	1.00	7.32%	15.28%
2023-02-10 11:25:00	98.38	1.00	7.26%	15.27%
2023-02-10 11:20:00	98.01	1.00	7.38%	15.25%
2023-02-10 11:15:00	96.54	1.00	7.33%	15.22%
2023-02-10 11:10:00	88.37	1.00	7.01%	15.20%
2023-02-10 11:05:00	86.03	1.00	6.83%	15.20%
2023-02-10 11:00:00	86.57	1.00	7.34%	15.21%
2023-02-10 10:55:00	86.72	1.00	6.83%	15.22%
2023-02-10 10:50:00	87.73	1.00	6.95%	15.22%
2023-02-10 10:45:00	93.19	1.00	7.02%	15.22%
2023-02-10 10:40:00	98.92	1.00	7.07%	15.20%
2023-02-10 10:35:00	98.89	1.00	6.96%	15.18%

SLO大盘-下半部分 对应指标的实时曲线，
明确不稳定的点具体信息



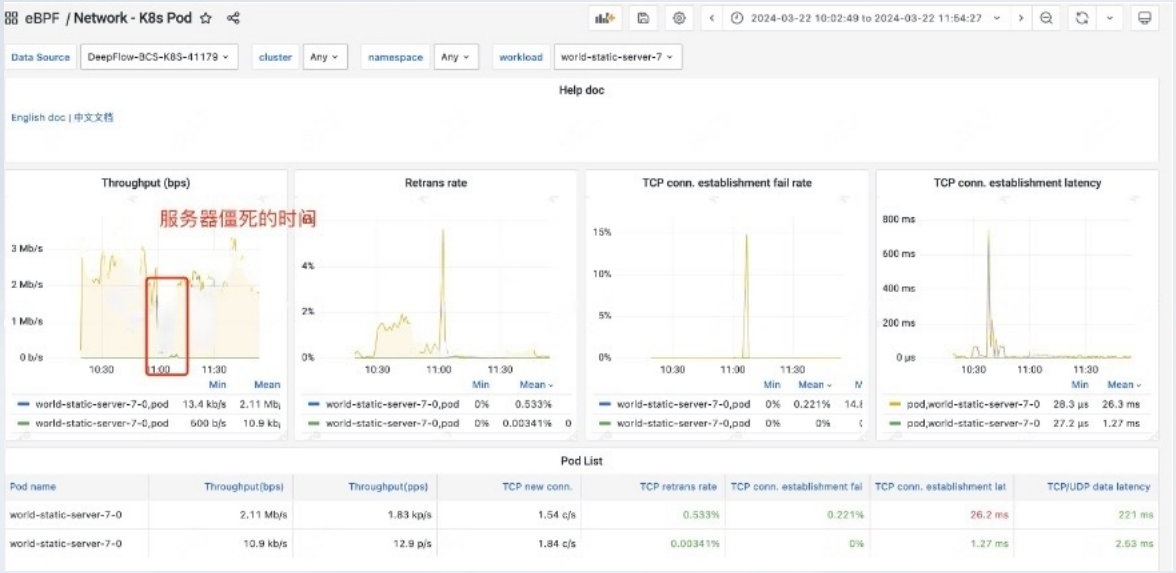


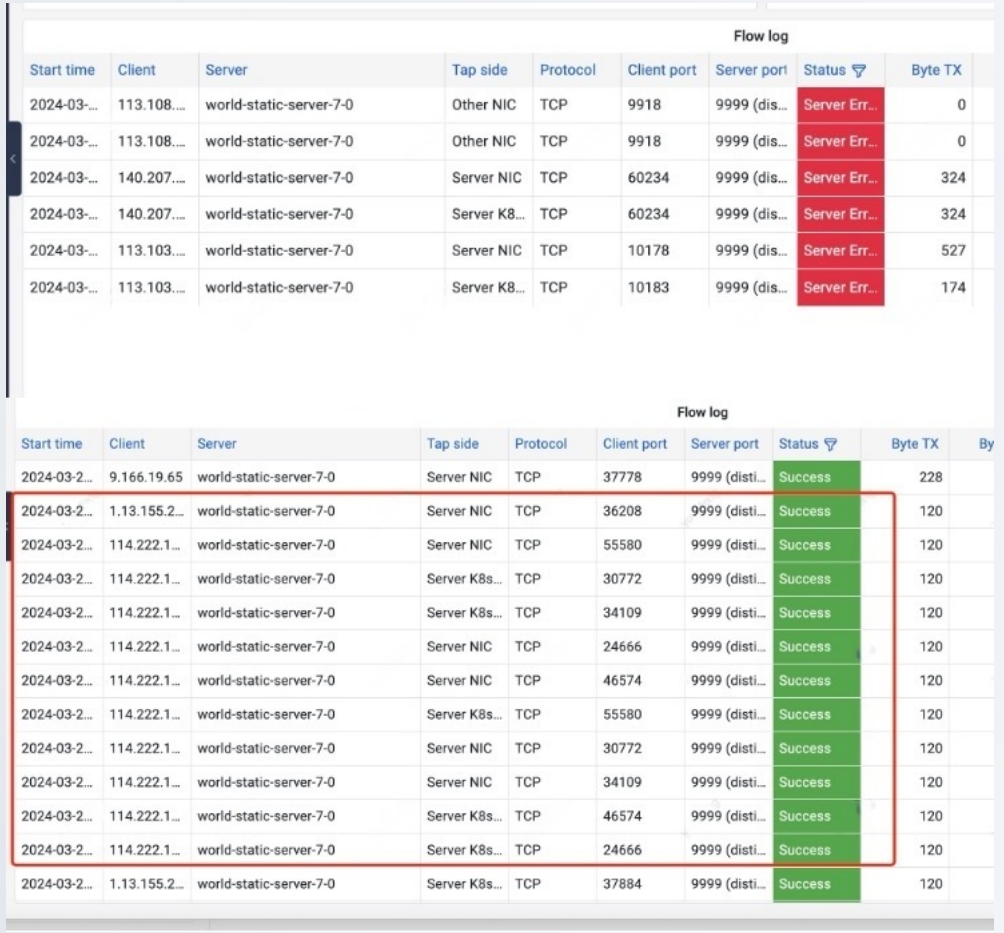
案例4：使用eBPF可观测，快速发现服务器端程序问题

SREelite

突然表现：在某个副本中进行战斗的玩家突然集体卡住

- 确定服务器是否在运行中：正常运行
- 确定服务器出入流量是否正常：流量掉0





- 从玩家到服务器的业务流量全部是server Error状态
- Clb到服务器的健康检查流量全部是success状态
- 是不是可以说明，只有涉及到业务协议的请求包才会出问题？

案例5：使用eBPF可观测，快速定位异常流量问题根因

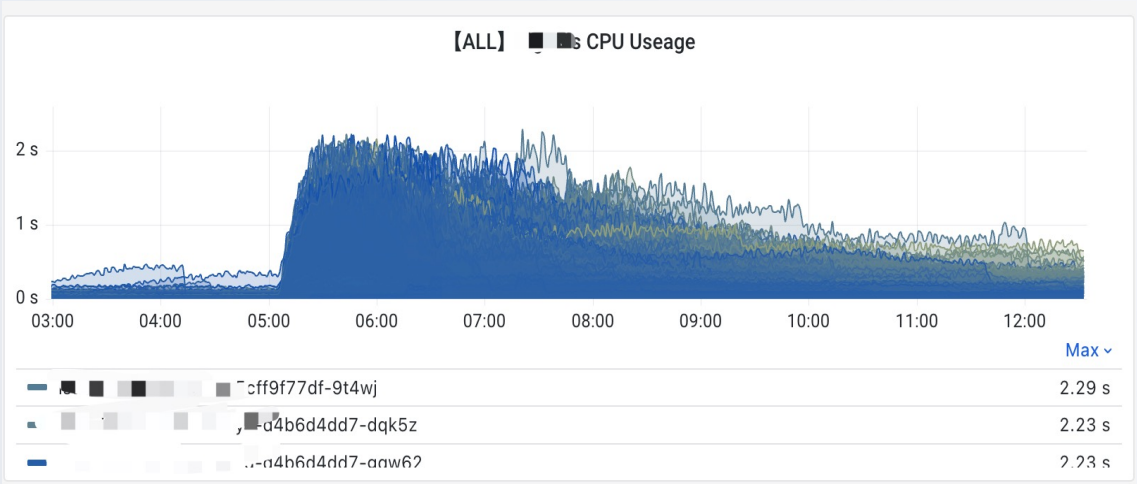
SREelite

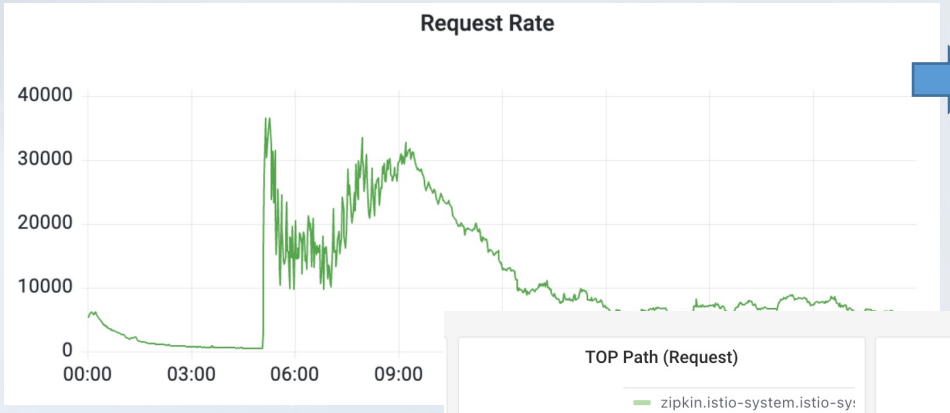
突然表现：整个链路每个组件的 cpu 使用率过高

- 怀疑玩家数增长：相较同期无明显增长
- 怀疑服务器产生过多 gc：不符合整个链路均增长的表现

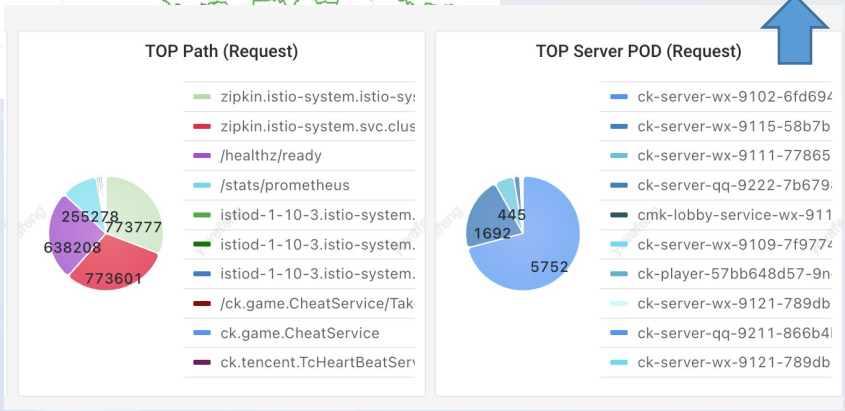
  整个链路均增长，说明导致问题的原因与网络有关

- 在玩家总数没有增长的情况下，是否有可能是单个玩家发起的请求数增多？





- 通过 request rate 确认，确实存在过高的 qps
- 通过 top path，top server pod，定位到 某pod 的接口异常非常多



- 反馈给研发，核实日志，很快定位到了根因

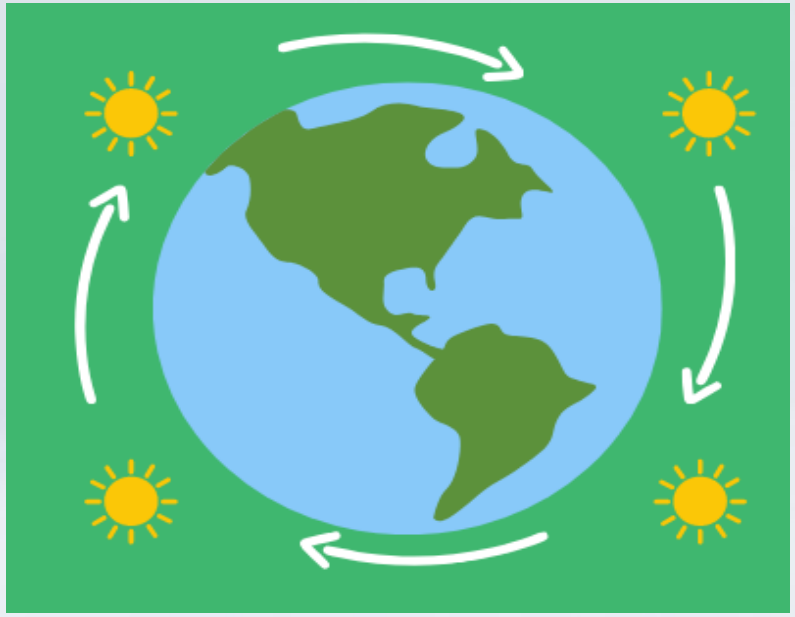
03

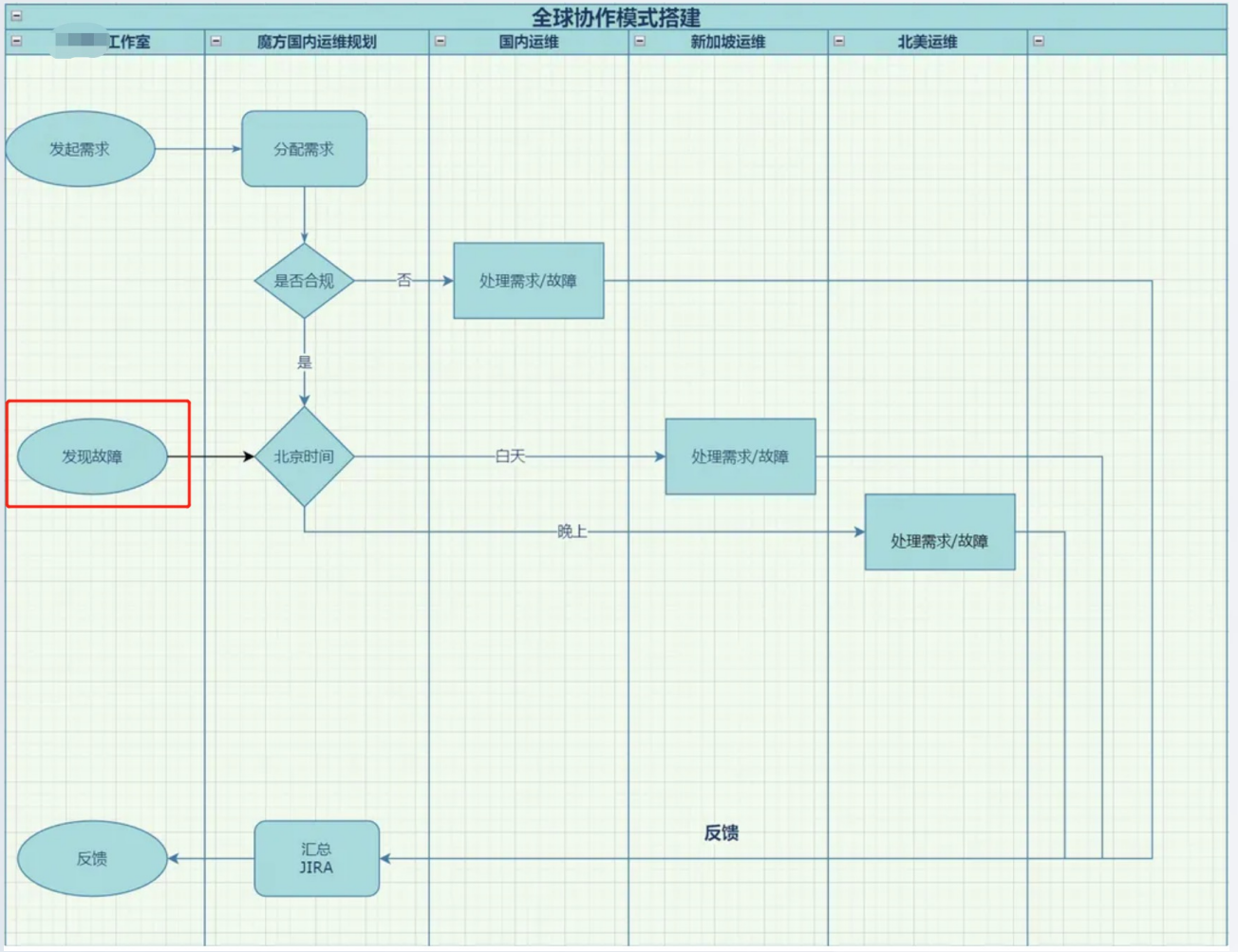
应对之道：快速响应并处理故障

建立ONCALL模式，保证故障得到快速响应

SREelite

全球化协作流程，保证业务能得到7*24小时全天候服务





建立有效的分级告警机制，减少告警信息洪流

SREelite

问题

每天收到N多告警，等于没告警

收到告警，该谁接收和处理?

大多业务噪声降低 90%+

如何解决告警风暴问题:

✓ 区分告警等级，使用不同告警策略

✓ 告警责任人自动识别，关注点分离

告警等级	判断依据	ONCALL	告警渠道	告警主要责任人
致命	生产环境为主：严重影响游戏用户体验，收入类等问题。比如大面积网络和组件类故障，关键模块机器重启，程序严重BUG类	SG（8：00-21：00） NA（21:00-8:00）	企微致命告警群+电话告警+短信+邮件	基础设施：SRE 程序类　：对应模块开发 公共数据类：SRE和开发
预警	问题堆积后，可能会产生致命问题，比如生产环境容量，性能类告警	SG（8：00-21：00） NA（21:00-8:00）	企微预警群+短信	基础设施：SRE 程序类　：对应模块开发 公共数据类：SRE和开发
提醒	信息事件为主，比如日志错误统计类，非生产环境告警	SG（8：00-21：00） NA（21:00-8:00）	企微提醒群	基础设施：SRE 程序类　：对应模块开发 公共数据类：SRE和开发
过滤	无效信息，不告警			

建立故障应急处理分工协助流程-Google的故障指挥体系

Incident Commander
故障指挥官
简称IC

IC，整个指挥体系的核心，职责是组织和协调，而不是执行，下面所有的角色都要接受他的指令并严格执行。

Communication Lead
沟通引导
简称CL

CL，负责对内和对外的信息收集及通报，这个角色一般相对固定，由技术支持、QA 或者是某个 SRE 来承担都可以，但是要求沟通表达能力要比较好。

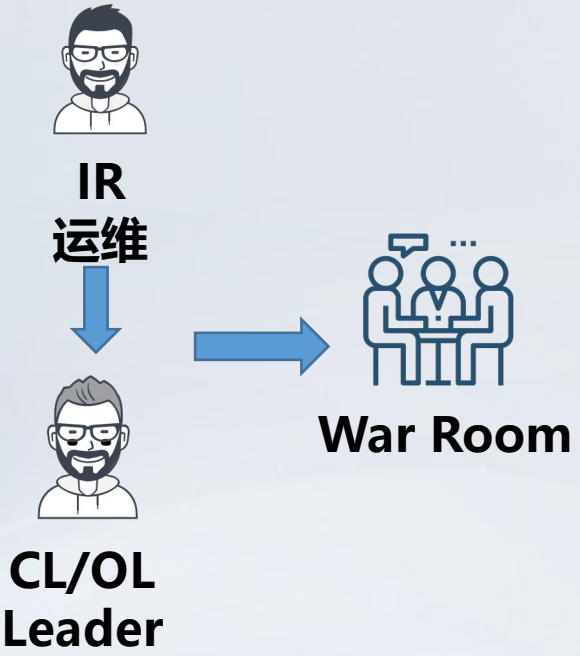
Operations Lead
运维指挥
简称OL

OL，负责指挥或指导各种故障预案的执行和业务恢复。

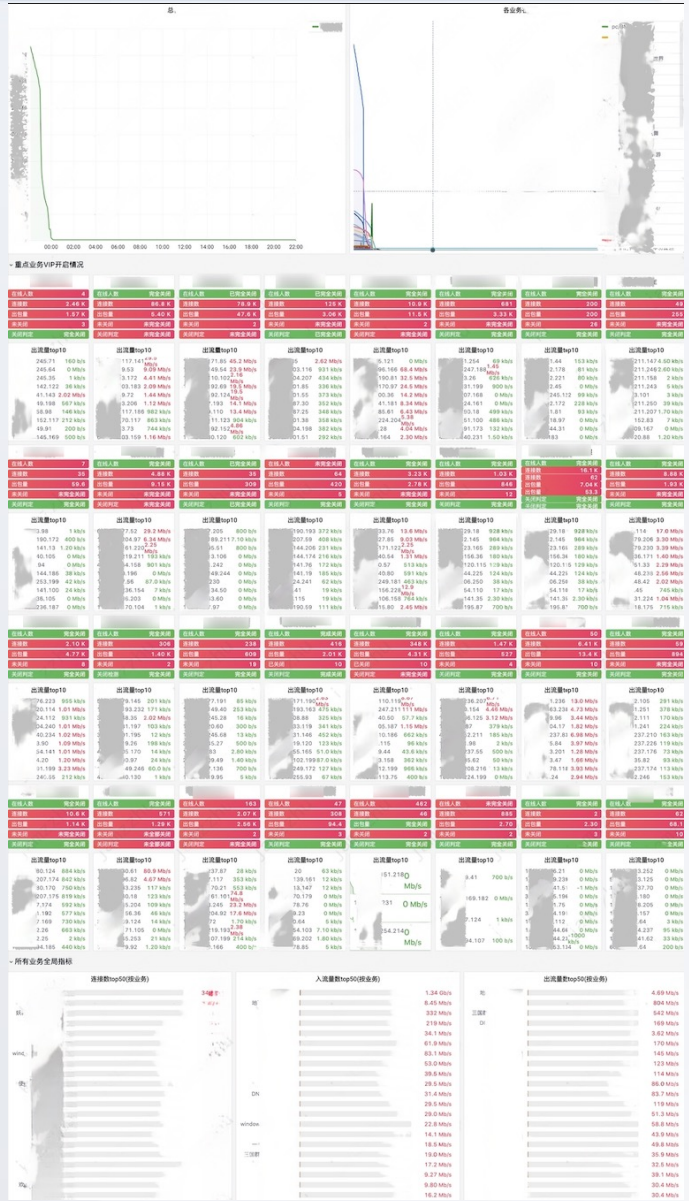
Incident Responders
故障处理
简称IR

IR，所有需要参与到故障处理中的各类人员，真正的故障定位和业务恢复都是他们来完成的，比如具体执行的 SRE、网络和系统运维、业务开发、平台开发、网络或系统运维、DBA，甚至是 QA 等

建立属于腾讯游戏的故障指挥体系



War Room成员组成				
角色	IC	CL	OL	IR
负责岗位	总监	Leader/运营规划	运营规划	各系统开发/运维/SRE



从实践经验来看，如果是大范围的故障，一般就是总监来承担IC职责，接下来他就可以从更高的层面组织和协调资源投入并有效协作。

这时运维回归到OL的职责上，负责组织和协调具体的执行恢复操作的动作。

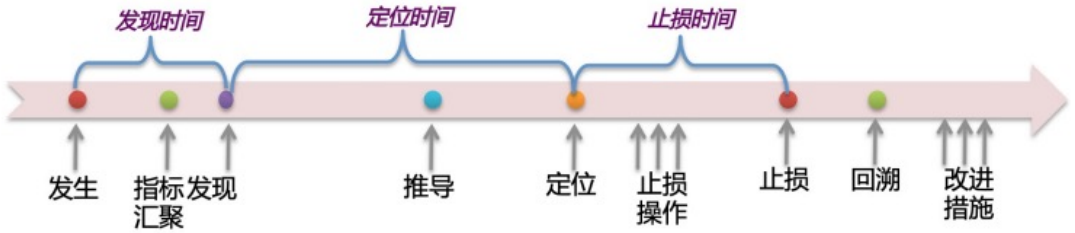
利用蓝鲸故障自愈快速恢复业务故障

SREelite

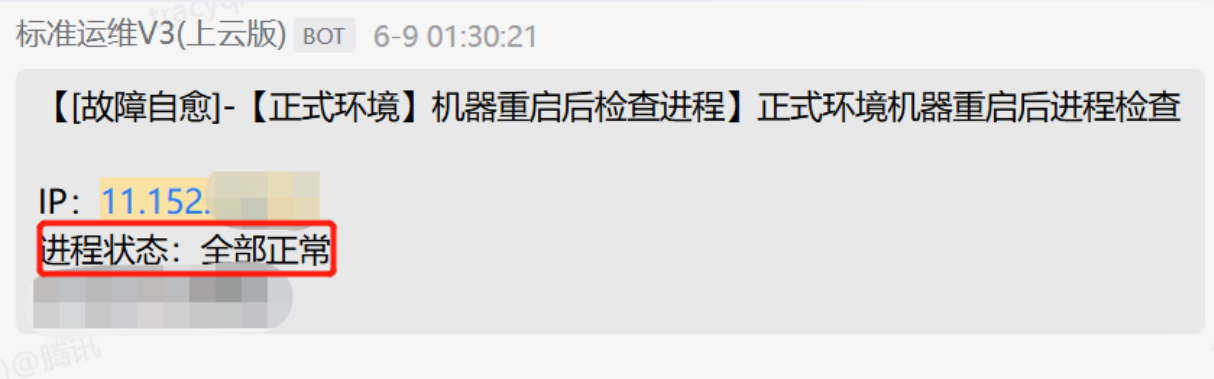
某业务使用故障自愈恢复业务进程　**1分钟**自动处理，减少MTTR时间

ID	套餐名称	套餐类型	关联策略	触发次数(近 7 天)	最近更新人
#107001	【标准运维】【现网】屏蔽集群扩容告警策略	标准运维	1	10	
#99592	【标准运维】【现网】zonesvr通知在线人数	标准运维	1	–	
#90275	123	一键拉群	–	–	
#67588	正式环境致命告警ITSM建单	流程服务	198	–	
#69188	【标准运维】【现网】故障自愈拉起单主机进程	标准运维	2	2	
#69163	【标准运维】【预发布】故障自愈拉起单主机进程	标准运维	1	–	
#65697	【通知套餐】【运维】xwork捕获CVM运行异常通知	标准运维	1	–	
#63429	【快捷套餐】转移主机至空闲机模块	标准运维	–	–	
#63187	【快捷套餐】重启网管Agent	标准运维	–	–	
#63186	【TCM】重启idipsvr和midas_ntf_svr（指定IP地址…	标准运维	–	–	

平均故障恢复时间: MTTR（mean time to restoration），指所有故障的平均恢复时间。



故障恢复，及时通知



利用蓝鲸故障自愈快速扩缩容

SREelite

处理套餐说明： 通过告警策略可以触发处理套餐，处理套餐可以与周边系统打通完成复杂的功能，甚至是达到自愈的目的。

ID	套餐名称	套餐类型	关联策略	触发次数(近 7 天)
#16410	故障自愈-全球战斗服自动扩缩容	标准运维	1	--
#9305	【不删档】【非DS Error】gdb coredu...	标准运维	1	--
#6286	【不删档】【非DS Error】corefile日志...	标准运维	2	--
#6285	【不删档】【tbuspp】corefile日志内容...	标准运维	2	--
#3992	【不删档】【DS Error】corefile日志内...	标准运维	2	1
#5806	【不删档】【自愈套餐】发送 CPU 使...	标准运维	3	--
#4204	【测试环境】corefile通知	标准运维	3	--
#3923	【CBT2】corefile日志内容通知	标准运维	--	--

每天定时跑检查结果

记录扩缩容，节省成本直接展示

自动扩缩容实际执行结果

时间	执行状态	大区	实际扩缩容数量	执行时常	缩减成本/月/天
2024-05-10 15:46:00	成功	弗吉尼亚	-2	3 hour	¥-2
2024-05-10 15:47:00	成功	弗吉尼亚	-2	3 hour	¥-2
2024-05-10 15:48:00	成功	弗吉尼亚	-2	3 hour	¥-2
2024-05-10 15:49:00	成功	弗吉尼亚	-10	3 hour	¥-1

04

复盘总结：从故障中学习并改进

SREelite

什么是故障复盘

简单重复10000次，不如有效复盘1次
把经验转化为能力
将失败转化为“有意义的失败”

东方：复盘
西方：结构性反思

故障复盘底层逻辑

- 故障是无法完全避免的
- 杜绝重复犯错，包容失败
- 根因分析要全面

在故障复盘中，发生故障后分析原因，总结经验教训，并制定改进措施，以防止类似故障再次发生。

故障复盘的目标

SREelite

找出故障原因	降低故障率	提升团队技能	提升客户满意度
硬件问题	充足预案	预防风险意识	及时响应和沟通
软件问题	故障演练	团队协作能力	问题解决和修复
系统问题	优化系统	问题解决技能	透明诚信
网络问题	提升软件质量	持续改进意识	预防和改进措施
配置问题	提升监控告警质量	上游思维	关注和重视
安全问题	提升系统稳定性	定期复盘	决心和毅力
人为问题	提升系统可靠性	知识库沉淀	……
……	……	……	

故障复盘关键要素

SREelite

尊重复盘 → **建好气氛** → **基于事实** → **识别故障**

- 不要零散时间
- 准备好会议
- 准备好材料和会议议题
- 邀请相关人员
- 会议纪要
- 总结和反馈

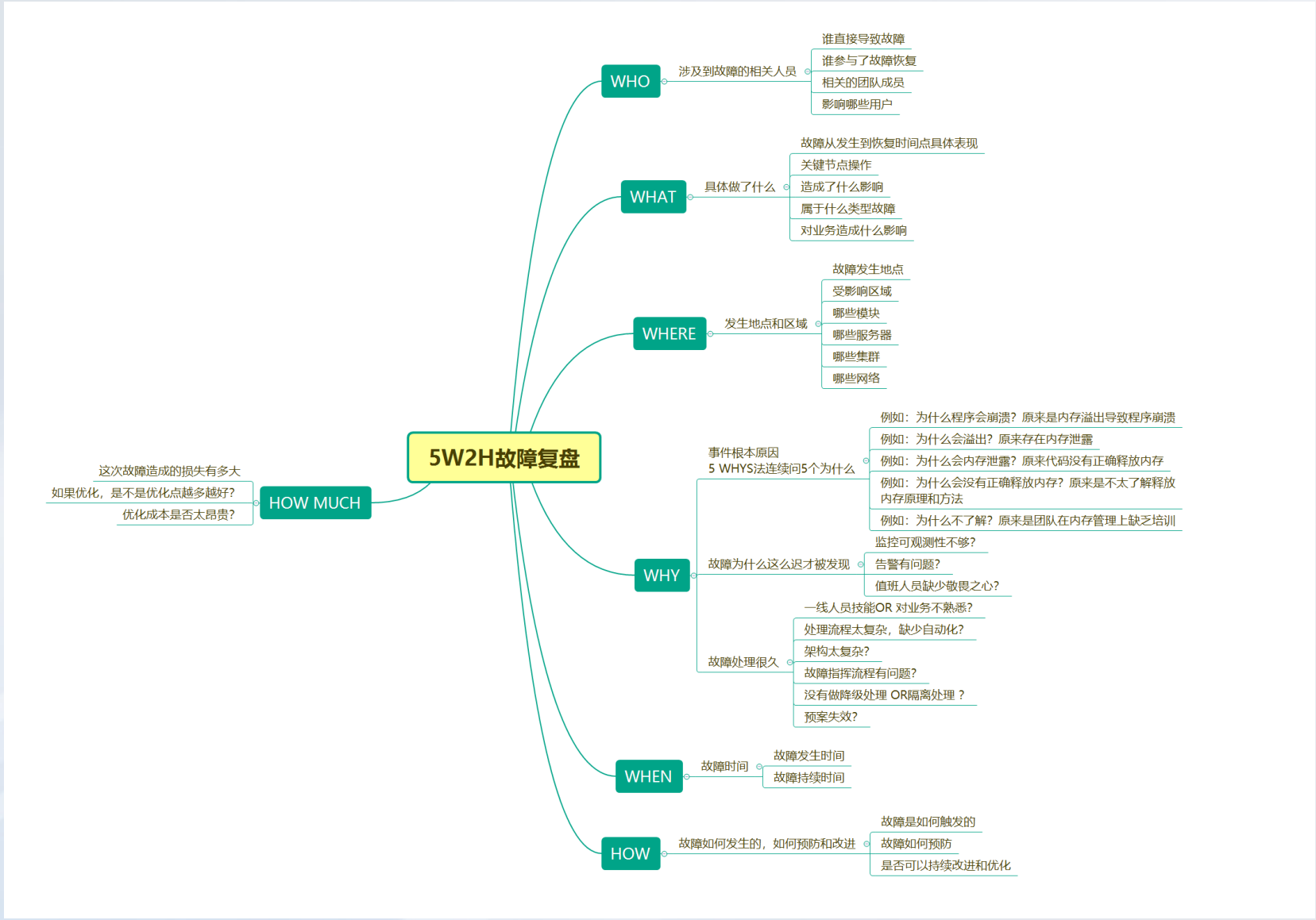
- 不以指责为目的
- 不要定责
- 积极献言，鼓励分享
- 开放和尊重
- 有效沟通

- 收集数据
- 分析数据
- 讨论和解释
- 基于事实的决策

- 识别故障
- 记录故障关键事件
- 分析故障原因
- 制定改进措施
- 沉淀知识库，分享

复盘方法

5W2H法


5W2H故障复盘
WHO
涉及到故障的相关人员
谁直接导致故障
谁参与了故障恢复
相关的团队成员
影响哪些用户
WHAT
具体做了什么
故障从发生到恢复时间点具体表现
关键节点操作
造成了什么影响
属于什么类型故障
对业务造成什么影响
WHERE
发生地点和区域
故障发生地点
受影响区域
哪些模块
哪些服务器
哪些集群
哪些网络
WHY
事件根本原因
5 WHYS法连续问5个为什么
例如：为什么程序会崩溃？原来是内存溢出导致程序崩溃
例如：为什么会溢出？原来存在内存泄露
例如：为什么会内存泄露？原来代码没有正确释放内存
例如：为什么会没有正确释放内存？原来是不太了解释放内存原理和方法
例如：为什么不了解？原来是团队在内存管理上缺乏培训
故障为什么这么迟才被发现
监控可观测性不够?
告警有问题?
值班人员缺少敬畏之心?
故障处理很久
一线人员技能OR 对业务不熟悉?
处理流程太复杂，缺少自动化?
架构太复杂?
故障指挥流程有问题?
没有做降级处理 OR隔离处理 ?
预案失效?
WHEN
故障时间
故障发生时间
故障持续时间
HOW
故障如何发生的，如何预防和改进
故障是如何触发的
故障如何预防
是否可以持续改进和优化
HOW MUCH
这次故障造成的损失有多大
如果优化，是不是优化点越多越好?
优化成本是否太昂贵?

复盘模板参考

SREelite

5W2H
根本原因分析（RCA）

某故障复盘模板

故障概述

1. 主题：〔简述什么时间点发生了什么动作，做了什么操作，造成什么影响〕
2. 业务：xx业务
3. 时间：xx时间-xx时间
4. 责任归属：xx
5. 事故影响：xx

故障原因

〔分析故障的直接原因和根本原因，包括人为因素、权限因素等，有针对性得预防类似事故再次发生〕

1. 平台权限控制xxx问题
2. 安全操作流程执行不严格
3. xxx

事件回顾-时间线描述

〔详细描述事件从需求产生、发生事故、处理过程、故障恢复的整个流程〕

1. xxx时间点，收到XX告警
2. XXX时间点，收到XXX告警
3. XXX时间点，运维开始处理故障
4. XXX时间点，确认是XX模块故障
5. xxx时间点，运维重启该模块进程，故障恢复

故障影响

〔对事故过程中项目影响与损失进行评估，了解事故严重程度〕

1. 影响小区掉线xxx人
2. 影响收入XXX

整改措施

1. 针对事故发生的原因提出整改措施，完善制度

总结与反思

1. 对事故反思与总结，分析事故中的问题和不足，举一反三，提出改善建议

Q&A

SREelite

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

SREelite

2003年
Google 成立 SRE 团队

2016年
孙宇聪翻译出版首部 SRE 著作

2020年
赵成牵头成立 SRE研讨社群

SREelite
2023年11 月
SRE 精英联盟正式成立

《SRE 实践白皮书》首次正式发布

2024年2月

联盟首次直播

2024年2月29 日

联盟首次线下 Meetup

2024年4月13 日

https://sre-elite.com

《SRE实践白皮书》

- 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。

关注我们

SREelite

公众号

视频号

B 站

YouTube