应急事件处理

应急事件处理

故障定级

- 事故等级制度

紧急事件响应

On-Call 机制基础

正确地On-Call

- On-Call轮值期间，作为生产系统的监管者，SRE思考和解决问题的方法论对正确地处理问题至关重要
- 为避免应急事故处理过程中On-Call工程师在压力下但凭直觉的操作，应该事先建立起规范、清晰地解决...

紧急事故的流程管理要素

紧急事故的处理流程机制示例

本节大纲

紧急事件响应

紧急事故管理

故障排查

事后总结

on-call机制

应急事件处理

系统可靠性是MTTF和MTTR的函数，评价一个团队将系统恢复到正常情况的最有效指标，就是MTTR

■任何需要人工操作的事情都只会延长恢复时间

■具有自动恢复能力的系统，即使有更多的故障发生，也要比事事都需要人工干预的系统可靠性更高

■必须人工介入的情况下，基于记录有最佳方法的事先预案进行处理能有效降低MTTR

切实降低MTTR中的人力介入时长的方式有两个

■资深、万能的工程师

■手持“运维宝典”且训练有素的On-Call工程师

- ◆“运维宝典”详细提供了各类事故的清晰处理机制
- ◆On-Call工程师周期性参与演练各类事故的处理
- ◆组织建设有清晰的故障定级标准和处理流程

故障处理过程中的效率取决于三个因素

■技术层面的故障隔离手段是否完备

■故障处理过程中的指挥体系是否完善，角色分工是否明确

■故障处理机制是否经过充分演练

故障定级

事故等级制度

团队只有知道相关问题的严重程度，才可以按照问题等级投入资源进行优化

	P0	P1	P2
影响时间长度	大于30分钟	大于15分钟	大于1分钟
功能比例	错误率大于50%	错误率大于20%，小于50%	错误率小于20%
资金损失	10万元	1万~10万元	0~1万元
舆论影响	客户投诉大于1000例	客户投诉大于200例	客服投诉大于10例

■从SRE团队的角度看，发生一个严重事故时，往往意味着线上业务的稳定性受到了很大冲击

■这个时候SRE团队会决策接下来的线上运维策略执行收紧原则，开始严格控制变更频率，加强变更评审等方式压制可能继续导致线上稳定性波动的因素

异常响应流程

异常处理环节的设计要求：从SRE的角度来看，异常处理流程应该固化到平台上

处理环节	要 求
异常发现	监控发现为主，用户反馈为辅
问题跟进	问题响应时间小于5分钟，一个人处理，一个人同步信息
问题升级机制	按照时间长度对问题进行汇报升级（15分钟升级通知到总监）
问题分析	需要完全定位问题
问题通知	有明确的通知机制，5分钟播报一次进度
问题解决	完全修复问题
问题后续追踪	暴露问题分析定性，安排跟进，形成处理闭环

紧急事件响应

东西早晚会坏的，“系统正常，只是该系统无数异常情况下的一种特例。

合理、周全的监控体系，是及时发现故障的根本前提

紧急事件响应

■及时发现故障、合理判定故障等级，尽早投入匹配的力量和资源介入并尽快完成恢复

■谁来负责响应及启动必要的响应升级：On-Call工程师

◆On-Call机制是确保服务稳定性目标的一种有效工具

◆On-Call工程师负责根据SLO、错误预算及故障等级评判标准决定要采取的应对措施

■如何处理故障

◆正确的方式未必是找到根因并一次性修复问题，而是应该先尽最大可能地让系统恢复服务，可用的应急措施包括故障转移（切换路由至可用的其它区域）、限流、降级和熔断等

◆缓解系统问题是第一要务，故障定位和排除则是次要目标

■故障定位和排除

◆掌握和熟练完成故障排查的通用过程

◆理解发生故障的系统的设计方式和构建原理

服务恢复后，要做事后总结并进行根因分析，从失败中学习和长进

■事后总结：记录事故详情，找出根因，并采取有效措施降低问题重现的概率，甚至避免其重现

■跟踪故障：系统性地从过去发生过的所有问题中总结经验教训

On-Call 机制基础

MTTI（从发现故障到响应故障）环节主要有两个任务

■判断出现的问题是不是故障，以及故障的等级

◆根据问题的危害程度（故障等级），判定需要投入的资源

◆对于大型问题，必要时可以联系其他团队，或者升级请求支援

■确定由谁来监管生产系统

◆On-Call轮值工程师

◆负责处理生产环境中即将或者正在发生的业务事故，以及评审对生产系统的变更请求

On-Call工程师负责确认告警信息、及时定位问题、并尝试解决问题，以更好地保障服务的可靠性和可用性

■在IT行业里，一般由专门的运维团队成员轮值

■轮值期间，需要在分钟级别执行生产系统的维护需求

◆面向最终消费者的服务，或者时间非常紧迫的服务，要在5分钟内响应

◆非敏感业务通常宽限至30分钟内响应

On-Call 机制（2）

On-Call准则

■非紧急的生产系统事件，例如低优先级的告警处理，或者新软件的发版可由on-call工程师在工作时间内评审或者执行

■生产告警信息的处理是第一紧急要务，它几乎超过一切其他活动，包括研发项目的进行

On-Call工程师角色的“高可用和负载均衡”

■可以在团队中同时配置主on-call者和副on-call者

■“高可用”on-call：副on-call者在主on-call者没有响应的情况下，作为备用对紧急事件进行响应

■“负载均衡”on-call：主on-call者负责处理生产系统中的紧急情况，副on-call者负责处理其他非紧急的生产环境变更需求

正确地On-Call

On-Call轮值期间，作为生产系统的监管者，SRE思考和解决问题的方法论对正确地处理问题至关重要

■现代理论研究表明，面临挑战时，人们会主动或非主动地选择如下两种应对方法中的一种

◆依赖直觉，自动化、快速行动

◆理性、专注、有意识地进行认知类活动

■处理复杂的系统问题时，第二种方式能以更周全的执行过程生成更好的处理结果

为避免应急事故处理过程中On-Call工程师在压力下但凭直觉的操作，应该事先建立起规范、清晰地解决问题的步骤，确保On-Call者可以冷静地审视和验证提出的所有假设，平稳地化解风险

■清晰定义的应急事件处理步骤并经常演练

■清晰的问题升级路线；在面临类似如下问题时，可采用的应急事务处理流程

◆处理复杂问题，需要同时引入多个团队时

◆经过一段时间的处理，仍不能解决问题时

■无指责，对事不对人的文化氛围

再次强调，SRE团队必须在大型应急事件发生之后书写事后报告，详细记录所有事件发生的时间线

紧急事故的流程管理要素

嵌套式职责分离

■分工明确，职责清晰

■手头任务过多，可以申请更多的人力资源，相应的管理职责由自己完成或转移给其他负责人控制

■紧急事故管理流程中的参与者

◆**事故总负责人**：组建团队、分配任务、协调资源

◆**运维指挥官和事故处理团队**：指挥官负责指挥事故处理团队执行具体的操作来尝试解决问题

◆**发言人**：向事故处理团队和所有关心本次事故者发送周期性通知，并维护事故相关的文档

◆**规划负责人**：为事务处理团队提供支持，处理一些待续性的工作，例如编写Bug报告、安排职责交接及各种后勤事务

控制中心

■建立作战室（war room）：将处理问题的全部成员集中办公，或通过电视电话会议的方式建立虚拟的作战室

■经常参与事故管理的人员也可以组建起单独的沟通渠道，例如即时联络群

实时事故状态文档

■事故总负责人最重要的职责之一就是维护一个实时事故文档

■该文档要允许多人同时编辑，以便于随时生成和汇总处理进度详情

明确、公开的职责交接

紧急事故的处理流程机制示例

故障发现后，On-Call工程师，是一开始时的事故总负责人，他有权召集相应的业务开发或其他必要的人员，快速组织作战室

■监控系统通常应该是用于发现故障的最重要依仗，用户反馈及舆情监测手段做为补充

若问题及恢复路线非常明确，则事故总负责人不变，他负责继续指挥、调度完成后续的处理流程，以恢复业务为最高优先级

■On-Call工程师借助监控系统及个人经验判断故障的大致原因，若能在短时间内恢复则立即执行恢复预案，否则，要立即启动应急升级预案

若问题复杂且发现影响范围较大，On-Call工程师可以请求更高级别的主管介入，例如SRE主管或总监等，并将总指挥权转移给高级别的负责人，On-Call则扮演运维指挥官的角色

■On-Call工程师启动作战室，编写简单的故障说明，同时开始召集各路相关人员

■人员就位后，汇报情况，并移交总指挥权

■此时的总体原则仍然是优先恢复业务，同时需要“发言人”准备要发布的公告

待生产基本恢复后，开始定位根因，并解决根因相关的问题

最后复盘故障，并更新“运维宝典”