



事件 ONCALL 中心建设方法

一站式处理值班 OnCall, 智能降噪

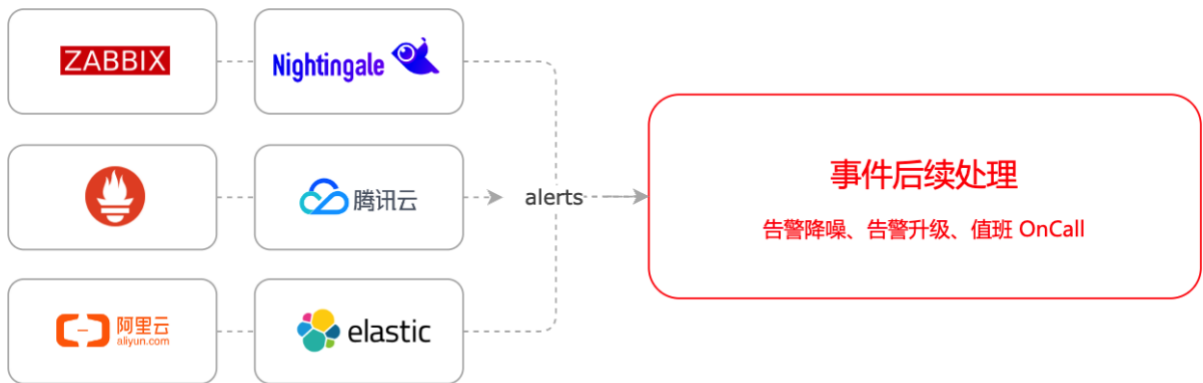


前言

市面上有众多监控系统，刨去商业软件不说，开源的就有 Nagios、Zabbix、Open-Falcon、Nightingale、Grafana、Prometheus、Elastalert 等等，还有云厂商提供的监控系统，比如华为云的云监控、腾讯云的云监控、阿里云的云监控，甚至有些云厂商会提供多个割裂的监控系统，比如阿里云不但有云监控，还有 ARMS，还有 SLS。

大部分公司都不会只使用一套监控系统，网络设备的监控可能采用的 Zabbix，Kubernetes 的监控可能用的 Prometheus（Kubernetes 可能有多套，以至于 Prometheus 可能有多套）或者 Nightingale，日志的监控可能用的 Elastalert，如果上云了，可能还会有多套不同的云监控（尤其是多云场景下）。

监控系统的重心，通常是采集、存储、可视化、生成告警事件，但通常都不具有完备的事件后续处理能力。这里说的后续处理主要包括：多渠道分级通知、告警静默、抑制、收敛聚合、降噪、排班、认领升级、协同闭环处理等等。监控系统或多或少都有一些这方面的能力，但是通常都不完备，而这，正是 [PagerDuty FlashDuty](#) 这种产品存在的价值。这些产品都是以 Duty 命名，核心就是支持告警 OnCall 值班处理的场景。



对于告警事件的后续处理，有哪些问题和需求以及何为最佳实践？我们从思路方法和工具实践两个方面分别进行探讨，下面先行探讨思路方法，看看要解决这些问题和需求，我们有哪些可能的解法。

思路方法篇

告警事件的后续处理：多渠道分级通知、告警静默、抑制、收敛聚合、降噪、排班、认领升级、协同闭环处理等等。看起来需求很多，最核心的痛点有两个：

- 告警太多，打扰太多
- 告警疏漏，无法闭环

我们先来看第一个痛点，首先分析一下造成告警太多、打扰太多的原因是什么，然后针对原因提出对应的方案。

告警太多的常见原因

最常见的原因，是**告警规则设置得不合理**。比如很多规则触发了告警之后，实际没有后续动作，只是起到常态化通知的效果，不需要排查，也不需要止损，甚至连个长线的 TODO 都没有。这类告警多了人就累了，当重要的告警来临的时候，也容易忽略。这样的规则如果不经过治理，日积月累，就会产生很多无用的告警。

第二个常见的原因是**底层出问题导致所有的上层依赖都告警**，越是底层影响越大，比如基础网络如果出问题，发出几万条告警都是正常的。

第三个原因是**渠道错配**。一些不重要的告警也使用打扰性很高的渠道发出，用户可能会觉得单一渠道不可靠，想用多个渠道同时发送的方式来保障告警触达率，这也属于告警规则配置不合理的范畴。

第四个原因是**预期内的维护动作**导致的。比如程序升级变更，如果进程重启时间过长，可能会导致关联的服务告警，或者某个机器重启，忘记提前屏蔽了，也会产生一堆关联告警。

了解了常见原因，下面我们来看一下有哪些常见解法。

优化告警规则

类似 [PagerDuty](#) [FlashDuty](#) 这种产品，一定程度上是可以解决一些告警过多的问题，但如果能从告警规则的源头做好优化，自然是事半功倍。很多公司的告警规则配置没有原则可循，每次故障复盘先看告警是否漏报，一线工程师为了不背锅，自然是尽量多地提高告警覆盖面，但这么做的后果，就是告警过多，无效告警占多数，长此以往，工程师疲惫不堪。

那么告警规则的配置应该遵照一个什么原则呢？虽然每个公司业务不同，总有一些通用的原则可循吧？的确如此，这里我分享一下我个人的做法，希望你有所启发。

每个规则都应该对应具体的 Runbook

Runbook 就是告警处理手册，也就是告警触发之后，应该细化排查哪些方面，按照一个什么方式执行动作，应该有一个手册参考。如果告警发生之后没有后续动作，那这个告警的意义就不大了。在 Nightingale 的告警规则配置页面，可以看到一个专门的 Runbook 配置，Grafana 的告警配置页面，也有一个 Runbook 的选项，就能看出他们对它的重视程度。

这个原则看起来是不是很合理？但是真要落地的时候，又会发现紧急需要处理的告警事件通常容易对应 Runbook，但是有些告警规则产生的告警确实没有那么紧急，有些只是想作为一个通知，好像又确实难以对应一个固定的 Runbook。

针对这两种情况，我的做法是：不紧急的告警，也必须要有动作，虽然这个动作可能不是立马执行处理，但至少创建个低优先级的工单之类的，或者提高告警阈值，等问题严重一些再告警。对于只是想通知一下的告警，其实都不算告警，只能看作是一种另类的报表和巡检手段，这样的“告警”就按照报表和巡检的逻辑来处理，比如把这类“告警”发到一个单独的邮件组或者单独的聊天群组，平时都不用关注，只要每天早上上班或晚上下班之前稍微看一眼就行，这样就可以减少打扰。

制定了这个原则之后，如果大家不遵守怎么办呢？还是有很多告警没有对应的 Runbook，作为管理人员，我们应该怎么处理？我的建议是分产品线统计一个指标：“Runbook 预置率”，就是各个产品线有多少告警规则配置了 Runbook，有多少没有配置，这个比例要统计出来，然后做成红黑榜，让大家去治

理，治理一段时间之后有经验了，知道预置率大概在一个什么范围是合理的，然后就可以要求大家至少达到预置率下限的值。否则，就一定是有问题的。

Runbook 这个配置原则，是我最为推荐的原则，效果非常明显，其次就是告警分级原则。

每个告警都应该合理分级

基本每个监控系统都支持为告警规则配置不同的级别，基本上每个监控系统的用户也都知道应该做分级告警。但是具体怎么分级，却没有一个行业共识，大家各做各的。这里我也分享一下我的理解，你可以参考借鉴。

首先，不同级别的告警应该对应不同的处理逻辑，这样分级才有意义，比如通知渠道不同，通知范围不同，或者介入处理的人的范围不同，处理时效不同，如果某两个级别对应完全一样的处理逻辑，就可以合并成一个级别。

我的做法是把告警分成 3 个级别。

级别	通知渠道	说明
Critical	电话、短信、即时消息、邮件	影响收入的、影响客户的，必须立刻处理
Warning	短信、即时消息、邮件	无需立刻处理，但是如果不处理，时间久了就会演化为 Critical 的问题，可以先放入 TODO 列表，手头上的紧急事务搞定之后就去处理
Info	邮件	每天下班前稍微看一眼，偶尔一两天忘了看也无伤大雅

另外，如果 Critical 的告警规则很多，大概率也有问题，说明系统架构不够鲁棒，出点什么事都要立刻介入，系统没有自愈能力。这样的系统，需要配备更多运维人员，而且还很难跟老板讲清楚价值。怎么办？这就需要制定运维准入规则，哪个系统要交给运维人员来运维，首先要提供一些信息。

- 相关联系人，出了问题能够及时找到人，联系不上得能直接联系研发领导。

- 服务相关信息，比如代码仓库、系统架构、依赖哪些服务、依赖哪些系统参数、哪些 JVM 参数、常见问题还有处理办法等等。

然后进行准入评审及准入测试，如果系统架构有明显问题，就没办法通过准入要求，不接受运维，如果老板要求必须接，那就只能加入了，或者明确说明在架构调整好之前，不负责 SLA，反推业务改造。

上面介绍的两个告警规则优化原则，是最重要的两个原则。照做的话，可以搞定大部分无效告警。

除了原则方面，另一个应对过多告警的方法就是靠产品工具了，比如告警事件在哪些时间段发送、如何过滤、如何屏蔽、如何抑制等等，通常，监控系统和统一的 OnCall 中心（[PagerDuty](#) [FlashDuty](#) 这种产品）在这些功能上会有一些的重叠，不过监控系统在这方面做得参差不齐，整体能力偏弱，使用统一的 OnCall 中心功能更强大，我们留待工具实践篇再详细阐述。

接下来我们聊一下“告警疏漏、无法闭环”的问题，核心就是告警发出来得有人处理，所谓的闭环，就是指告警发出、认领、协作处理、问题恢复、复盘改进的整个过程。

虽然事件降噪的几个手段落实之后，事件数量确实变少了，但是处理告警事件显然不是一个让人愉快的事情，不愉快的事情就要团队共担，所以第一个手段就是排班，专人做专事。

排班，专人做专事

这个手段听起来并不高无上，但确实非常有效。值班期间虽然提心吊胆的，生怕背锅，但因为是轮班制，心里总有个盼头，挺过这个周期就好了。

轮班的人在值班期间是第一责任人，会拿出 120% 的精力来处理问题，责任到人显然更容易推进问题解决，其他不值班的人则可以心无旁骛地做一些长线的事情，不至于总是被告警打断。

排班系统通常不开源，通常是作为事件中心的一个功能，PagerDuty 就提供了排班能力，即使没有系统支持，也建议人为制定一个排班表，把这个制度落实下去，对告警闭环处理也会有很大帮助。

值班人员在值班期间，虽然已经高度重视了，但也难免疏漏，这就需要告警升级机制了。

告警升级机制

告警升级是指在第一责任人收到告警之后没有及时响应，然后系统自动通知二线、三线人员的一种机制。一线人员没有及时响应的原因可能有很多，比如手机静音了没有听到，晚上睡着了，或者临时出去有事忘带手机了等等。这个时候系统发现某个告警一直没有恢复，也没有被认领，一段时间之后，就应该通知值班人员的领导或者二线备份人员，如果二线人员也迟迟没有响应，就应该继续往上升级。

告警升级机制需要认领功能的配合，也就是一线人员收到告警之后要通过某种机制告诉系统：“我已知晓告警，现在我开始处理了，你不要升级了”。典型的认领功能一般是做在页面上的，告警后打开告警事件管理中心，选中相关告警一键认领，也可以通过上行短信或即时通讯工具中的上行回调机制来完成。

升级机制会给值班人员很大的压力，毕竟谁也不想稍不留神就把电话打到老板那里，所以一般只有严重的告警才会启用升级机制，警告或者通知性质的告警都不用启用升级机制。当然，这个规范怎么定，各个团队可以自行商定。

通过排班、认领、升级这些机制，可以确保告警递达指定的人，但要处理告警的话，只有值班人员自己就未必搞得定了，需要有协同机制把相关人都拉进来一起处理才可以。对于某个故障，可能同时有多个告警事件产生，大家基于一个统一的故障协同，而不是基于一堆事件分别协同，这就需要把这多个事件收敛成一个故障，下面我们来聊一下这个收敛逻辑。

告警收敛逻辑

一般收敛逻辑是三级收敛，event -> alert -> incident。举个例子，最原始的告警事件，比如 host1 在 timestamp1 产生了一条 cpu_usage_idle 的告警，我们称为一个 event。如果没有恢复，一段时间之后，比如 timestamp1 + 60min，一般会再发出一个告警，还是 host1，还是 cpu_usage_idle 这个指

标。很明显，这两个告警事件是有关联关系的，指代的是一个问题，只是时间戳不同，这样的两个 event，就可以收敛为一个 alert。

从实现上来说，告警策略（也称告警规则）+ 指标标签集的哈希值，可以作为 alert 的唯一标识。比如刚才的例子，告警策略的 ID 假设为 32，标签集是：["name=cpu_usage_idle", "host=host1"]，这两个时间戳产生的告警事件，哈希值都是一样的。

计算方法是：

```
hash(32 + ["_name_=cpu_usage_idle", "host=host1"])
```

从 event 到 alert 的这个收敛逻辑，我们叫做一级收敛。只有这个收敛逻辑还不够，告警信息还是比较散，不能基于这些散乱的告警分别做协同，把多个 alert 收敛成一个 incident（故障），基于 incident 做协同才比较方便。但是，event 到 alert 是有一个固定的收敛逻辑的，可以通过程序自动收敛，而 alert 到 incident 却很难自动收敛。不过业界也会有一些常见的做法，下面我举几个例子。

1、根据时间做收敛

把告警中心收到的所有告警，按照时间维度做收敛，比如按照分钟颗粒度，一分钟内所有告警收敛成一个故障，下一分钟所有告警收敛成另一个故障。显然，一个故障内的多个告警相互之间可能没有关联关系，所以这种收敛方法不是太好。

2、根据时间 + 标签做收敛

除了时间维度，再加上某个标签作为收敛维度，比如机器标签，某个时间段内所有 A 机器的告警收敛成一个故障，所有 B 机器的告警收敛成另一个故障。或者按照服务维度，某个时间段内所有 A 服务的告警收敛成一个故障，所有 B 服务的告警收敛成另一个故障。看起来效果好多了，只是没办法和现实中的告警和故障建立完美的对应关系，不过从降噪收敛角度来看，够用了。

3、根据时间 + 文本相似度做收敛

文本相似度需要引入算法，但是算法总得有个规律，我们很想把某个故障相关的告警聚拢到一起，但是显然，很难有个行之有效的规律，没有规律的算法效果自然好不到哪儿去。

既然没办法把告警自动收敛成故障，那就手工来做。一个故障关联的关键告警，还是相对容易区分的，只要把关键告警关联到故障，后续基于这个故障做协同就可以了。所谓协同，一个是信息同步、协同处理，一个是共同复盘、管理跟进项。

故障协同处理

首先，并不是所有的告警都需要升级成故障协同处理。一般来讲，如果告警可以被值班人员直接处理掉，对别的团队负责的服务没有影响，不需要通知别的团队，通常是不需要升级成故障的，在告警层面来协同就可以了，自己团队内部消化掉；如果值班人员和他所在的团队没办法独自处理告警，才需要升级成故障，拉其他团队的人进来一起处理。

多个团队共同处理一个故障，不同团队的人会发现一些不同的线索，需要及时同步给所有相关的人，这个时候就可以在故障下面添加评论，其他人就可以及时看到。等到止损之后，大家还要根据故障时间线复盘，产出一系列跟进项，这个时候就需要这个故障管理模块具备跟进项管理的功能，或者至少能够跟任务管理系统良好打通。

有了这样一个故障协同的机制之后，故障被处理掉的概率就大幅提升了，后续再配合一些运营统计手段，统计各个团队的平均故障止损时间，建立红黑榜，大家就会有更高的热情来处理故障。当然，人的热情再高，也不如机器来得快，如果有些告警能够直接关联自动化处理逻辑，无疑可以大大增加事件闭环率。

告警自动处理

很多监控系统都可以配置 Webhook，当告警触发之后自动回调某个 HTTP 接口，来串联一些自动化的逻辑，让告警事件无人值守自动处理。比如某个机房的某个服务挂掉了，Webhook 的逻辑是自动调用切流的接口，把服务流量切走，这样来达到止损的目的。

告警自动处理的这段逻辑，未必一定能够做到告警自愈，有的时候只是使用这个机制来抓现场，也是非常有价值的。比如某个进程挂掉了，在挂掉的时候我想知道当时机器的一些运行情况，比如各项资源的占用

情况、系统日志的信息等等，我们就可以借助告警自动处理的这个方式，来自动跑个脚本抓取当时机器上的一些现场信息，相比收到告警之后手工登录机器查看要高效得多。

如上，是从思路方法层面，对事件的处理做了逻辑讲解。要求所有的监控系统实现这些能力不太现实，而且会造成一个一个的事件孤岛，所以典型的做法是把所有监控系统生成的事件统一聚合到一个平台来处理，这就是 OnCall 中心，下面我们以 [FlashDuty](#) 来举例，讲解 OnCall 中心的工具实践。

工具实践篇

称手好用的工具是可以大幅提升效率的，同时，好的工具可以沉淀最佳实践，沉淀经验，假设由你来设计一款 OnCall 产品，处理告警分发相关的这一系列需求，你会如何设计呢？接下来，我们站在设计者的角度，来讲解产品设计逻辑和实践方法，会更容易理解。

空间管理

通常来讲，一个公司不但会使用多个监控系统，而且会有很多个团队，如果所有的告警事件都在一个地方查看、管理，就会相互打扰。所以，OnCall 中心首先要设计一个协作空间的概念，来归类处理不同的事件，比如根据团队划分，或者根据系统、子系统划分。FlashDuty 第一个菜单就是协作空间，就是这个设计初衷。



比如我们团队是负责公司的支付系统，我们就可以创建一个以“支付”命名的协作空间。之后把支付团队相关的告警都接入这个协作空间，支付团队可能用了 Zabbix、Prometheus 等多个监控系统，所以，OnCall 这个产品需要提供多种数据集成方式，让告警事件很方便地上报上来。

集成中心

比如要接入 Prometheus 的告警事件，就需要创建一个 Prometheus 类型的集成 (Integration)，要接入 Nightingale 的告警事件，就需要创建一个 Nightingale 类型的集成 (Integration)，点击上例中的“支付”协作空间，进入协作空间详情，其中有个【集成数据】的入口：

The screenshot shows the 'Integration Center' for the team '二期测试' (Secondary Testing). At the top, there are three summary cards for the past week's performance:

- 过去一周故障 MTTA: 1 小时, 环比上周-87.9%
- 过去一周故障 MTTR: 3 小时, 环比上周-70.06%
- 过去一周故障数量: 616 条, 环比上周-15.73%

Below these are navigation tabs: 故障列表, 集成数据 (selected), 分派策略, 降噪配置.

The main section is titled '专属集成' (Dedicated Integrations) and contains:

- '自定义事件' (Custom Events) with a gear icon and a right arrow.
- 'cloudwatch.alert' with a camera icon, a right arrow, and a warning message: '推送地址发生变化, 请及时更新'.
- '二期测试' (Secondary Testing) with a red cube icon and a right arrow.
- A dashed box with a plus sign and the text '+ 新增一个集成' (Add a new integration).

Below this is the '订阅规则' (Subscription Rules) section:

- 'ALL' rule with a minus sign icon, last modified on 2023-06-08 15:35 by '头铁科技kk'.
- A dashed box with a plus sign and the text '+ 新增一条规则' (Add a new rule).

Finally, the '排除规则' (Exclusion Rules) section:

- '21' rule with a checkmark icon, last modified on 2023-06-15 19:15 by '头铁科技kk'.
- A dashed box with a plus sign and the text '+ 新增一条规则' (Add a new rule).

上例中我已经创建过多个集成了，你的环境是新的，只需要点击【+新增一个集成】，选择集成类型，随便输入一个集成名称，就可以创建一个集成。

创建完了集成之后，点击这个集成查看详情，会看到一个专属的 URL，以及相关的接入文档，去监控系统里配置 Webhook，Webhook 地址就写这个 URL，这样监控系统告警的时候，就会把告警事件推给 FlashDuty 的这个“支付”协作空间了。

在协作空间下面创建的集成，我们称为「专属集成」，还有一类集成称为「全局集成」，在集成中心这个菜单下。监控系统通过 Webhook 发给「全局集成」的告警事件，怎么进入协作空间的呢？通过在协作空间里配置订阅规则。

订阅规则

订阅规则是一些过滤条件（通过事件标签、属性等），用于匹配告警事件，匹配到的告警事件，自动进入这个协作空间。



The screenshot shows the '集成数据' (Integration Data) tab in the FlashDuty interface. It is divided into three sections: '专属集成' (Exclusive Integrations), '订阅规则' (Subscription Rules), and '排除规则' (Exclusion Rules). The '订阅规则' section is highlighted with a red border.

- 专属集成 (Exclusive Integrations):** Contains two integrations: '自定义事件' (Custom Event) with the latest event time '2023-07-13 14:33:02', and '二期测试' (Phase 2 Test) with the latest event time '2023-07-08 10:42:40'. There is also a 'cloudwatch.alert' integration with a note to update the address. A '+ 新增一个集成' (Add a new integration) button is present.
- 订阅规则 (Subscription Rules):** Contains one rule named 'ALL' with the last modification time '2023-06-08 15:35' by user '头铁科技kk'. A '+ 新增一条规则' (Add a new rule) button is present.
- 排除规则 (Exclusion Rules):** Contains one rule with the number '21' and the last modification time '2023-06-15 19:15' by user '头铁科技kk'. A '+ 新增一条规则' (Add a new rule) button is present.

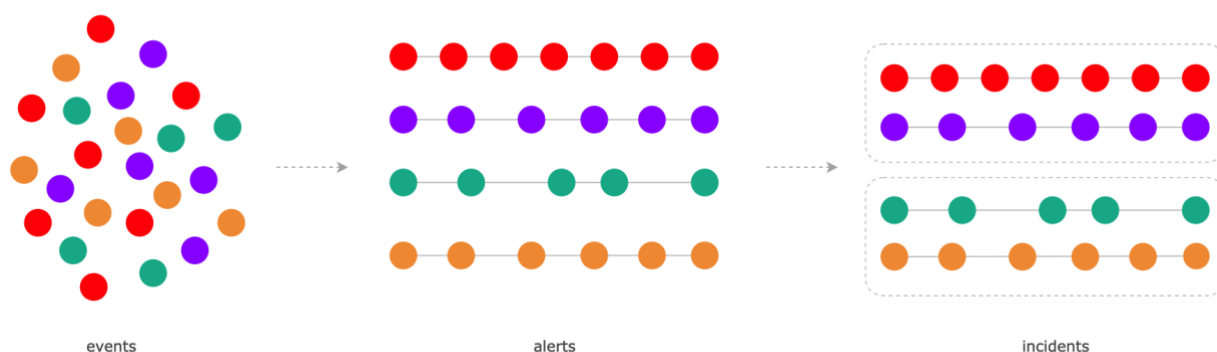
比如：根据标签 来过滤告警事件，把过滤到的告警事件订阅到“支付”协作空间。在“私有云”协作空间创建订阅规则，根据标签 来过滤告警事件，把过滤到的告警事件订阅到“私有云”协作空间。

排除规则

告警事件进入协作空间之后，有些特殊的告警事件想要丢弃掉，即可使用排除规则实现，配置位置就在订阅规则下面，这里不再赘述。

分派策略

告警事件进入 FlashDuty，会自动收敛，目前采用三级收敛：event -> alert -> incident，具体逻辑可以参考：【思路方法篇】-【告警收敛逻辑】章节。这样的三级收敛机制，会有非常好的降噪效果，大幅减少打扰。示意图如下：



监控系统会产生原始的告警事件 (event)，属于同一个告警的多个事件被合并成告警 (alert)，类似的告警 (比如某个标签相同，或者文本相似度很高) 被合并成故障 (incident)，最终通知用户的是一个故障，大幅降低了打扰性。

不同的告警事件，通常有不同的分发逻辑，比如不同时段不同的分发逻辑：白天用短信通知，晚上用电话通知，比如对象存储的告警要发给存储团队，物理机故障要发给运维团队。这都可以灵活定义。

2 策略配置

The screenshot displays a configuration interface for a strategy. It is divided into several sections:

- Time Range:** A section labeled "在" (In) with two tabs: "全部" (All) and "分段" (Segmented). Below the tabs is a "选择日期" (Select date) input field and a "开始时间 → 结束时间" (Start time → End time) range selector with a clock icon and a plus sign.
- Alert Conditions:** A section labeled "针对" (Target) with two tabs: "全部" (All) and "部分" (Partial). Below the tabs is a large container for conditions. It contains two input fields: "请选择或输入 Key" (Please select or enter Key) and "请先选择 Key" (Please select Key), separated by an "==" operator. Below these fields is a "+ 添加" (Add) button. To the right of this container is a "+ 新的条件" (New condition) button.
- Delay and Template:** A section labeled "延迟" (Delay) with a numeric input field set to "0" and the text "秒后" (seconds after), followed by "按照模版" (according to template). Below this is a dropdown menu set to "默认告警模板" (Default alert template) and a "进行通知" (Send notification) button.

也可以配置聚合窗口，比如延迟 120 秒，如果在延迟等待期内，告警自动恢复或被人工处理，则不会发送该条告警。

OK，接下来就是通知给谁以及如何通知的问题了。比如通知某个人或者通知某个团队，也可以通知某个值班表，值班表的值班人接收告警。不同的严重程度的告警，还可以有不同的通知媒介，比如飞书、钉钉、企微、电话、短信、邮箱等。

环节1
✕

每隔 分钟告警 1 次，最多告警 次。告警将通知到

📅 值班表

👤 团队 > 于双羽的团队
+

👤 个人

以单聊渠道通知 🔔 请确认您已完成设置

Critical 飞书:Feishu ✕

Warning 飞书:Feishu ✕

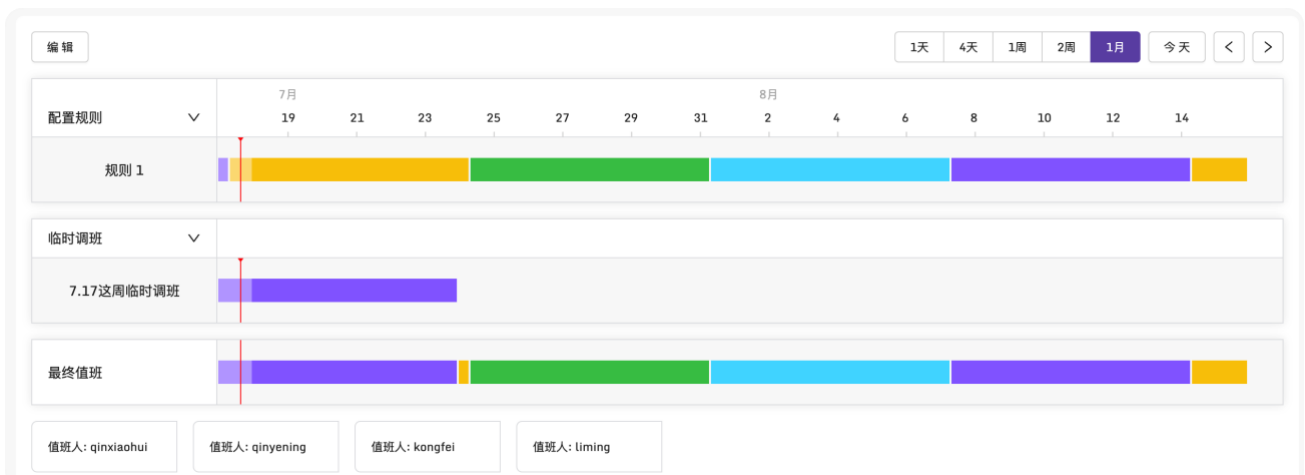
Info 飞书:Feishu ✕

以群聊渠道通知

超过 分钟后告警仍然未恢复且没有被人工处理，则升级到下一通知环节。

通知了之后，相关的人可能没注意到，可以配置重复通知，比如 10 分钟之后再次通知。如果多次通知，接收人一直没有响应，就要启动升级策略了，比如 30 分钟后，告警仍然没有恢复而且没有被人工处理，则升级到下一个通知环节（通知环节可以有多个）。

通知的触发，最为推荐的是值班表（OnCall 排班是践行 SRE 的有效手段），其次是团队（团队可以降低人员管理的负担），其次是个人（最不推荐，难以维护）。FlashDuty 提供了值班表功能，可以做日常排班以及节假日临时排班：



在 FlashDuty 里，通知的对象不是原始的告警事件，因为原始的告警事件可能会非常多，如上文所述，FlashDuty 会把事件聚合为告警，告警聚合为故障，最终通知的是故障。那具体如何聚合呢？

告警聚合

事件到告警的聚合比较容易，通常是用类似下面的算法来计算不同事件的关联关系：

```
hash(32 + ["__name__=cpu_usage_idle", "host=host1"])
```

这个值姑且称为事件 Hash，相同 Hash 的事件就被聚合为一条告警。更复杂的是告警到故障的合并，当前我们支持基于规则的聚合，后面会基于算法聚合：

告警聚合 ?

智能聚合 规则聚合 不聚合

i 模式说明 当告警的部分属性或标签与某个未关闭故障完全相同时，将告警合入故障。

聚合维度

支持设置多组条件，任一组条件匹配，即认为告警与故障相似；

每一组条件内，支持选择多个属性和标签，仅当所有属性和标签完全相同时，该组条件匹配。

labels.check



或

告警标题



新增

聚合窗口

未开启，将持续合入新告警，直到故障被关闭

风暴预警

聚合告警达到 条时，再次提醒我。

比如基于告警规则标题做聚合，某一时刻，基础网络故障，有 1000 台机器同时报了失联告警，就可以很方便地合并成一个故障，只通知这一个故障即可，大幅降噪减少通知打扰。

故障抖动

有的时候会出现一会告警一会恢复一会又告警一会又恢复的情况，称为告警抖动，FlashDuty 针对这种情况做了专项支持：

故障收敛

60 分钟内，相同的故障发生 2 次及以上则进入抖动状态，未来 120 分钟内不再发送新的故障通知

编辑

告警静默

静默规则通常用于预期内的维护行为。比如计划重启某个机器，那可以提前屏蔽这个机器相关的告警，避免操作的时候造成打扰。

2 静默时间

您指定的时间段内，告警将不会触发。

单次静默 周期静默

周一 × 周三 × 周四 × 周五 ×

02:00 → 05:00

选择日期

开始时间 → 结束时间

3 静默条件

在静默时间内，符合静默条件的告警将不会触发。您至少需要添加一项静默条件。

请选择或输入 Key == 请先选择 Key

+ 添加

+ 新的条件

屏蔽支持某个时间段内单次屏蔽，也支持周期屏蔽，比如固定的周末不发告警。

抑制规则

典型的场景是 Critical 的告警抑制同类的 Warning、Info 的告警。比如产生了两个告警事件，一个是 Info 事件：机器的内存使用率超过 70%，另一个是 Critical 事件：机器的内存使用率超过 90%，实际机器的内存使用率已经 95%，此时就只会发送 Critical 的事件，Info 的就被忽略了。

2 抑制条件

当新的告警满足以下条件，

集成	==	专属 自定义事件 ×	⊖
且	严重程度	==	Warning × Info ×
+ 添加			

+ 新的条件

且 24 小时内有满足以下条件的活跃告警 ②，

集成	==	专属 自定义事件 ×	⊖
且	严重程度	==	Critical ×
+ 添加			

+ 新的条件

且新的告警与活跃告警存在以下相同项，

属性	请选择
标签	resource ×

新的告警 将被抑制。

相关策略配置好之后，就可以收告警了，比如利用钉钉发送，会呈现为一张告警消息卡片：



Warning-订单系统

#8726BC cpu idle is too low

🕒 触发时间: 2023-07-21 17:45:10

⚠️ 聚合告警: 2条 (风暴中)

description :Cpu used too much (current 95%), please check the process or follow the runbook.

check : cpu.idle<10%

cluster : nj

metric : node_cpu_seconds_total

resource : es.nj.03

service : engine

跟进处理

直接关闭

临时屏蔽



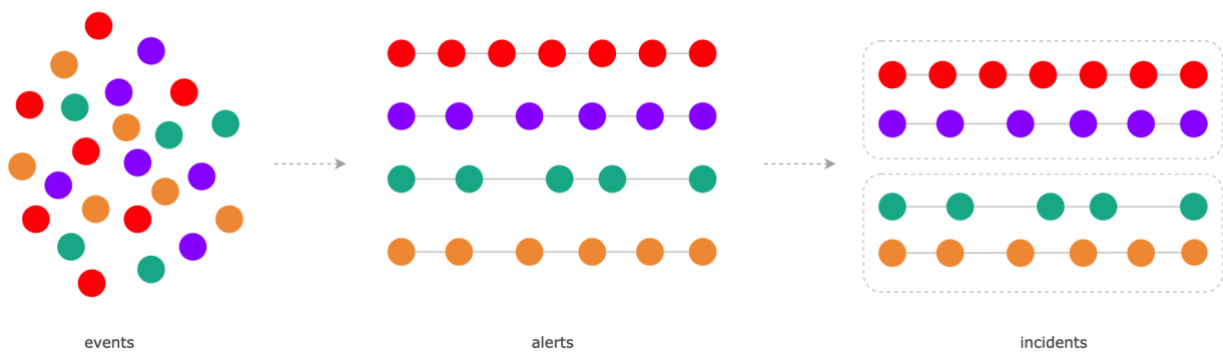
🔴 当前故障已聚合2条告警，触发告警风暴，请加急处理！

交互功能仅在账户关联后且license充足时可用 [去关联](#)

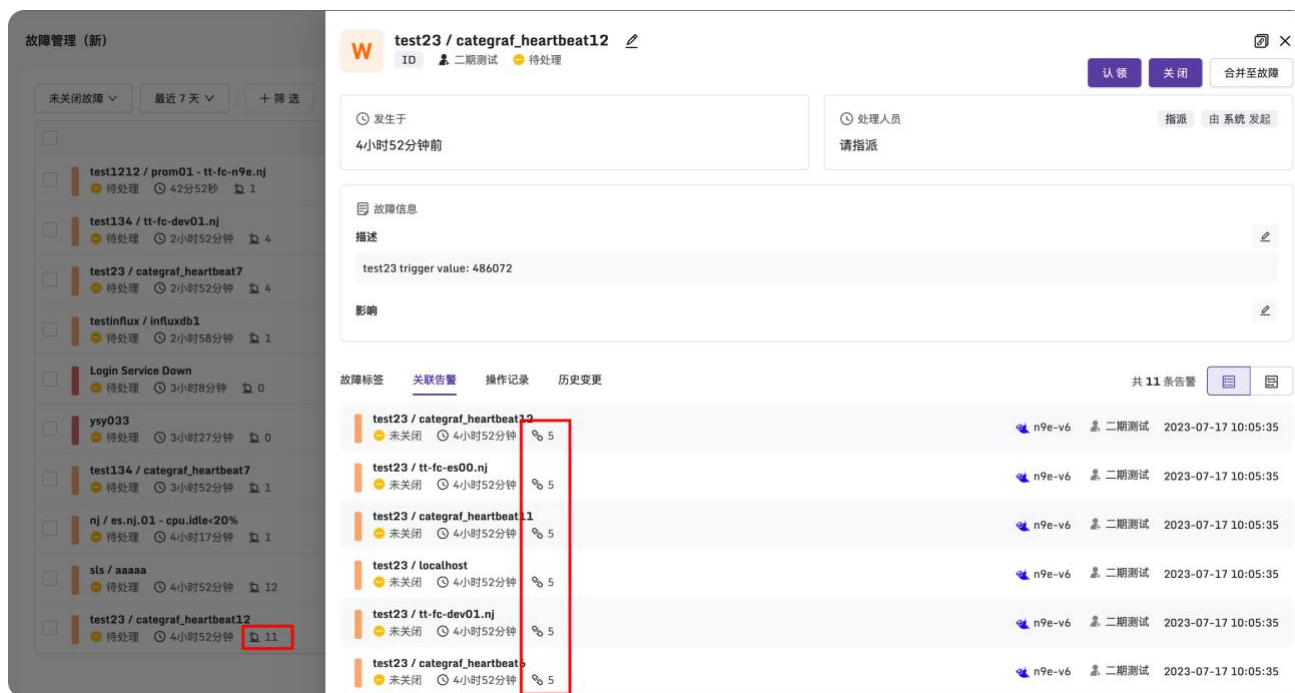
我们可以直接在钉钉（或飞书、企微等 IM）内部点击跟进，或直接关闭、临时屏蔽，方便地移动办公。当然，也可以登录 FlashDuty，在 WEB 上查看，需要有个非常直观的页面告诉用户您当前有哪些告警，分成了几类，每一类有多少条之类的。

告警/故障查看

OnCall 中心需要对接各类监控系统，同一时间可能会有很多告警事件发到 OnCall 中心，在中心统一查看、处理，这需要一个良好的组织形式，否则看起来就会很混乱。上文中我们介绍了两级收敛机制，events -> alerts -> incidents:



与其看到最底层的 events，我们肯定更希望看到 incidents，incidents 数量少，所以通常来讲，首先去故障管理里查看当前的故障（incidents）列表，每个故障关联了哪些告警，告警关联了哪些事件，也可以顺便查看。



如上图，最外层是故障列表（看起来比较多，这是我们的测试环境，正常来说，生产环境的故障不会很多，否则就说明出大问题了），每个故障关联了多少个告警，有个小警铃的图标可以看到，点击这个故障，右侧会出现一个抽屉，看到故障详情。关联告警这一栏，可以看到这个故障关联了哪些告警，每个告警也带有一个小锁链的图标，展示了关联的原始事件个数。

另外，FlashDuty 不但提供了故障视图，还直接提供了告警视图，两个视图都有两种展示方式：列表展示方式、聚合展示方式，方便您查看处理。

另外，FlashDuty 提供了和 IM（飞书、企微、钉钉等）深度集成，用户无需使用电脑，在手机上就可以快速查看故障/告警信息，比如在路上，赶去拿电脑的过程中，就可以提前快速了解相关信息，极大提升故障排查、止损效率。

告警/故障处理

通常，我们并不会基于告警来做协同，更多的是基于故障来做协同。点击某个故障，可以看到故障详情，会有认领、关闭、合并故障、评论等相关操作，示例图如下：

W test23 / categraf_heartbeat12 ID 二期测试 待处理 认领 关闭 合并至故障

发生于 5小时10分钟前 处理人员 请指派 指派 由系统发起

故障信息

描述 test23 trigger value: 486072

影响

故障标签 关联告警 操作记录 历史变更

输入评论内容，支持 markdown 语法 预览

评论

头铁科技kk 发起评论 14秒前 2023-07-17 15:15:38

国内到美东机房的网络堵塞了

系统 聚合告警达到 2 条, 触发风暴预警 5小时10分钟前 2023-07-17 10:05:52

系统 触发新故障 5小时10分钟前 2023-07-17 10:05:51

Warning test23 / categraf_heartbeat12

对于一些大故障，跨多个团队，拉齐信息是非常关键的，如果有某个团队发现了一些线索，可以通过评论的方式让其他团队快速知悉，新进的故障处理人员也可以通过这些评论以及故障关联的告警快速得知故障历史信息，快速启动排查工作。

下个版本还会继续增强和 IM 的联动，在 FlashDuty 中的一些评论回复，会自动发到 IM 端，进而提升协同效率。

另外，我们非常建议大家认真处理每一个故障，认真填写故障处理过程、止损手段，这是极好的知识库，未来再发生类似故障的时候，都可以快速参考。如前文所述，如果某个故障发生之后没有对应的处理动作，这个故障对应的告警规则可能就是不合理的，需要优化。

总结

告警事件的后续处理，不只是发往各个通知媒介那么简单。涉及到收敛、降噪、排班、认领、升级、协同、IM 打通 等非常多的细节功能，各个监控系统通常不会在这个方面发力，但是告警的统一处理又是非常强的需求，故而，我们推荐您使用 FlashDuty 来处理。下面是 FlashDuty 相关信息：

- 注册地址：https://console.flashcat.cloud/signup?from=wp_oncall 注册之后就两周全功能免费体验期
- 联系我们：<https://flashcat.cloud/contact/> 任何问题（采购、产品使用等）都可以联系我们