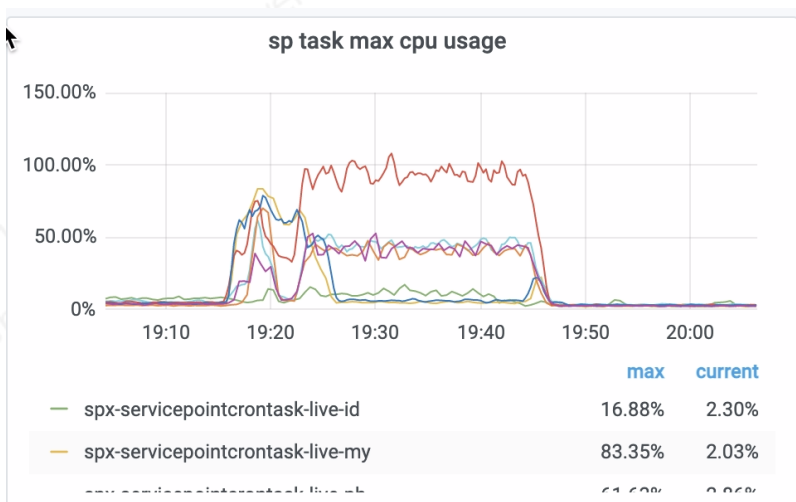


2022.04.07-SSC kafka集群宕机对spx的影响复盘

故障发生时间	发生时间 发现时间 恢复时间 持续时间	2022.04.06 19:14 2022.04.06 19:15 2022.04.06 20:41 87min
故障处理人	SRE：zepeng.yao，中间件：baomao.fang SPX：lianjie.chen、huan.chen、chao.chen、shaofeng.hu、谢磊、李川	
故障责任主体	/	
故障报告来源	Noc监控告警	
故障类别		
故障描述	SSC kafka集群宕机导致各条业务线作业受影响	
影响评估	业务无反馈，PS Team未报单，技术层面评估走异步不影响（轨迹没有自动重试）	
处理过程	1、2022.04.06 19:20 供应链监控i大群报kafka故障，其中有一台broker OOM问题，同时spx的fms应用开始报错；spx-trackserver 报错，rt增大；spx-servicepointapi rt增大，spx-servicepointcrontask CPU增长；  2、2022.04.06 19:22 Sre开始排查并处理kafka集群问题  3、2022.04.06 19:37，驿站spx-servicepointcrontask开始切换开关，绕过kafka执行异步任务；查看监控异步任务正常消费； 2022.04.06 19:46，spx-servicepointcrontask应用CPU立即下降至正常状态；	



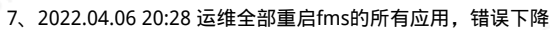
4、2022.04.06 19:42，运维重启，并切走了Leader。kafka集群恢复



5、2022.04.06 19:42 fms python的很多应用需要重启才能恢复



6、2022.04.06 19:20-19:42期间，spx-track-server持续报错，连接不上kafka集群



8、2022.04.07 少锋和Sre跟进kafka集群问题，确认是kafka集群的Leader在故障期间没有切换

Geoffrey Gao

20位成员

O406 kafka故障

1856860064

16:57

Calvin Wan · (call me if it's urgent) 1856860064
没有，具体的根因还在定位。@Geoffrey Gao (高锋) 13659215618 这个县简单写一个当前问题定位现状和我们现在可以做的事情吧。后续根因出来再一起复盘一下

ok, 收到，早上那个topic详情跟报告 我今天一并 输出一下

我们应用这边想确认一个点：
19：42分重启前，Leader有没切走？ 还是等19：42左右重启后才手动切换Leader。

因为期间，我们应用一直报“no leader”

这个信息很重要，我们要确定是客户端问题，还是集群问题。如果客户端问题，我们需要推升级

Geoffrey Gao (高锋) 13659215618

Hu Shaofeng (胡少锋)
我们应用这边想确认一个点：
19：42分重启前，Leader有没切走？ 还是等19：42左右重启后才手动切换Leader。
...

集群期间controller已经不在线了，所以会报错no leader

嗯，好的

原因分析	<ol style="list-style-type: none"> 1. kafka集群一台机器OOM，目前Sre还在复盘定位问题，还没给出原因 2. python应用不具备自动重连能力，需要重启应用，才能重连kafka 3. go应用依赖的chassis的kafka客户端sarama是v1.26.4版本，该版本存在一定风险，会导致producer和consumer存在重连失败的风险。详细参考 https://www.jianshu.com/p/f6e350c456fe
改进方案	<ul style="list-style-type: none"> • spx-servicepoint/smart-sorting的cpu高的排查--- Huan Chen，下周五之前反馈 • python的kafka客户端是否不具备重连能力排查--- Shaofeng Hu 找SRE一起跟进，下周五之前反馈 • 目前Br已经部署了kafka灾备方案，SEA+TW机房也需要按照计划陆续部署kafka灾备- lingfei.wang@shopee.com Shaofeng Hu Q2完成其他地区的部署 • 告警缺失，需要梳理相关告警项，配置告警（kafka集群的重要指标告警）。--- Shaofeng Hu 5.1前输出 • 轨迹补充自动重试逻辑-- chuan.li@shopee.com