云原生学院

云原生社区 Cloud Native Community

Chaos Mesh 在网易伏羲私有云自动化故障注入实践

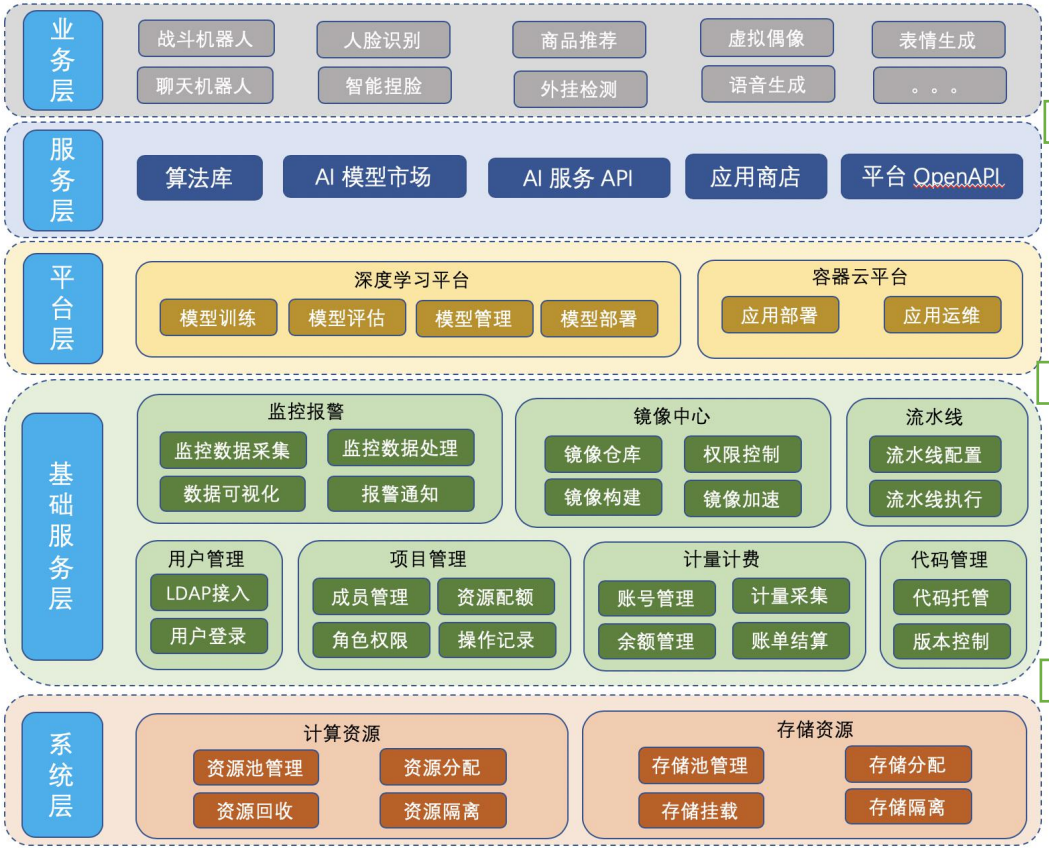
Speaker Name：张慧 网易伏羲

Speaker Title：网易伏羲私有云质量保障负责人、Chaos Mesh 布道师、云原生社区 Stability SIG 发起人

Email：zhangui05@corp.netease.com

网易伏羲私有云简介

网易伏羲



AI 模型

支撑游戏业务

云游戏

为什么混沌测试

一：问题现象描述

CMS监控到 HDFS NameNode备用节点意外终止，重启备用NameNode节点后CMS报告整个HDFS服务不可用

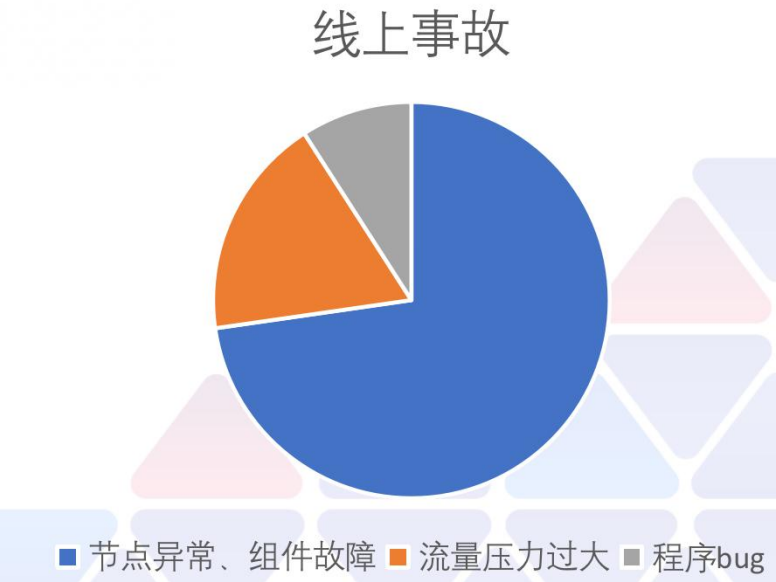
二：问题实际影响

1. HDFS服务不可用持续19分钟
2. 部分数据表的当天数据需要通过Kafka中的数据重建，无数据丢失

BUG3.0 #64417 【FU103】开发联调集群无法正常部署更新应用 ☆

描述

1. danlu-api 的 CPU 负载太高（400% 多）
2. 重启了 dl-scheduler 和 dl-istio 解决了该问题



为什么混沌测试

据外媒报道，亚马逊云端服务Amazon Web Services(AWS)25日遭遇了持续数小时的故障，导致部分网站和服务系统崩溃。

AWS的服务状态页面上的通知显示，因其处理大量数据流的服务器Kinesis出现问题，导致一些网站的“错误率增加”，亚马逊已经对该问题进行了修复，但完全恢复还需要一段时间，并贴出了当前受到影响的服务。

据AWS称，这次宕机仅影响亚马逊23个地理AWS区域之一，但这个问题已经严重到影响到了大量互联网公司的服务。

谷歌再次出现全球大宕机涉及YouTube、Gmail等一系列服务 ...

7 天前 — 刚刚，谷歌服务器又一次全球宕机！ 7月14日讯，据彭博社报道，Google公司服务器中断，导致旗下的YouTube、Gmail、Google Drive、Google ...

www.leiphone.com › news

崩溃！谷歌再次出现全球大宕机，涉及YouTube、Gmail等一 ...

7 天前 — 刚刚，谷歌服务器又一次全球宕机！ 12月14日讯，据彭博社报道，Google公司服务器中断，导致旗下的YouTube、Gmail、Google ...

www.pingwest.com › ...

Google宕机45分钟，全世界网友急疯了-品玩

7 天前 — 于是相互询问能够登陆Google服务成为了今天的流行问候语。事实上，美西时间的凌晨3:47分左右，Google全球宕机了。 Gmail，YouTube， ...

tech.163.com › ...

全面崩溃！Google服务又一次全球宕机|谷歌|gmail|youtube ...

理想下，系统用不宕机，100%可用

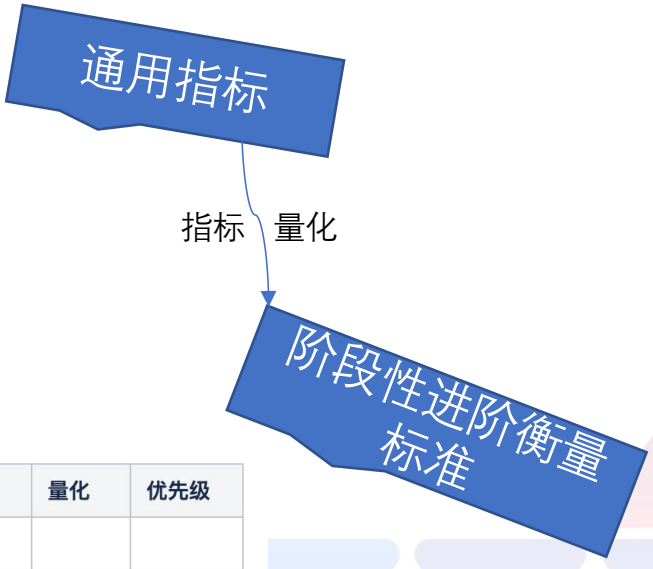
比如机房突然断电

事故突然的到来

为什么混沌测试

云原生社区 Cloud Native Community

SLA(service level agreement)指标				
	通俗叫法	可用性级别	年度宕机时间	每天宕机时间
可用	1个9	90%	36.5天	2.4小时
基本可用	2个9	99%	87.6小时	14分钟
较高可用性	3个9	99.9%	8.76小时	86秒
具有故障自动恢复能力的可用性	4个9	99.99%	52.6分钟	8.6秒
极高可用性	5个9	99.999%	5.25分钟	0.86秒

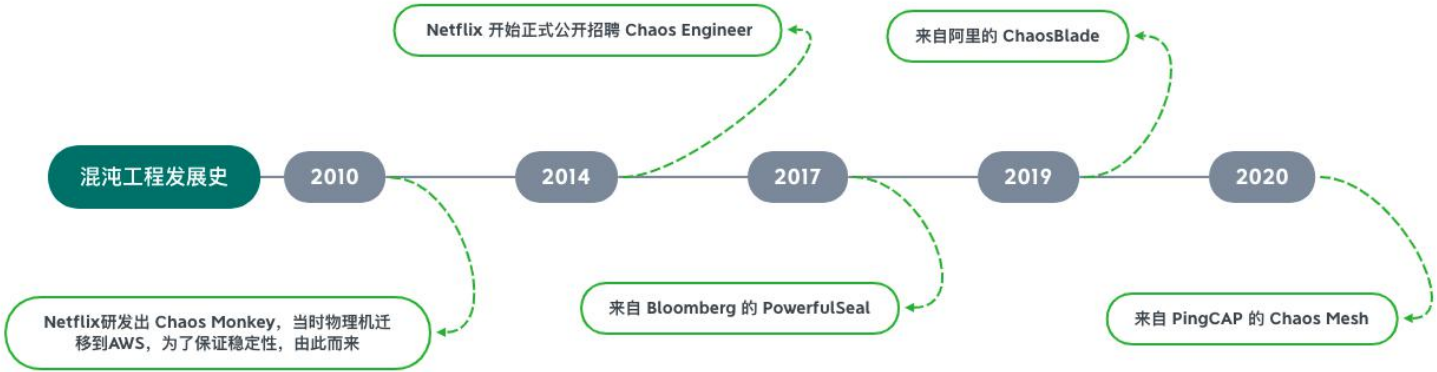


评价指标	具体指标	说明	量化	优先级
可用性	集群可用性	用来评估 k8s 集群在各类故障场景下的可用性		
	产品可用性	用来评估产品各组件以及平台在各类故障场景下的可用性		
	灾难恢复	用来评估特定灾难场景后的恢复情况，比如：降级、熔断		
稳定性	集群类稳定性	用来评估 k8s 集群本身的稳定性		
	产品类稳定性	用来评估产品各组件的稳定性		
	系统类稳定性	用来评估 kernel、docker 等系统底层等对产品稳定性的影响		

什么是混沌测试

混沌工程旨在将故障扼杀在襁褓之中，也就是在故障造成中断之前将它们识别出来。通过主动制造故障，测试系统在各种压力下的行为，识别并修复故障问题，避免造成严重后果。

混沌工程将预想的事情和实际发生的事情进行对比，通过“有意识搞破坏”来提升系统稳定性。



故障注入

鲁棒性

混沌工具

云原生社区 Cloud Native Community

Observability and Analysis - Chaos Engineering (8)

Chaos Mesh ★3,068
Cloud Native Computing Foundation (CNCF)
Funding: $3M

Chaos Toolkit ★1,152
ChaosIQ

Chaosblade ★3,340
Alibaba Cloud
MCap: $693.52B

chaoskube ★1,274
chaoskube

Gremlin
Gremlin
Funding: $28M

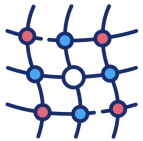
Litmus ★1,429
Cloud Native Computing Foundation (CNCF)
Funding: $3M

PowerfulSeal ★1,499
Bloomberg

steadybit
steadybit

混沌工具

混沌工具	Platform(s)	主机故障	容器故障	GUI	CLI	指标/报告	快速停止故障	目标随机化	社区活跃程度
Gremlin	SaaS	✓	✓	✓	✓	✓	✓	✓	
Chaos Monkey	Spinnaker	✓		✓				✓	
ChaosBlade	Docker, Kubernetes, bare metal, cloud	✓	✓		✓		✓		
Chaos Mesh	Kubernetes	✓	✓	✓	✓	✓	✓	✓	✓✓
Litmus	Kubernetes	✓	✓	✓	✓	✓	✓	✓	✓
ChaosToolkit	Docker, Kubernetes, bare metal, cloud	✓	✓		✓	✓		✓	
PowerfulSeal	Kubernetes	✓	✓		✓	✓			

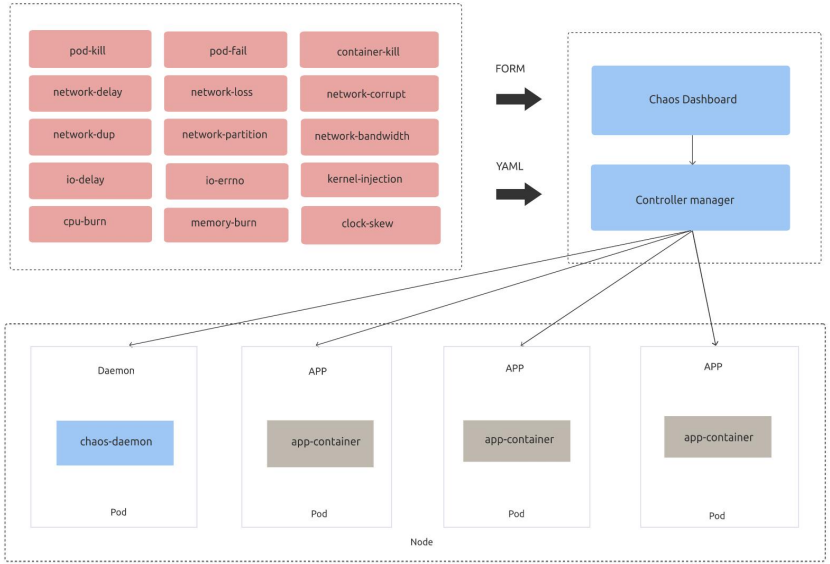


Chaos Mesh®

A Powerful Chaos Engineering Platform for Kubernetes

为什么是 Chaos Mesh

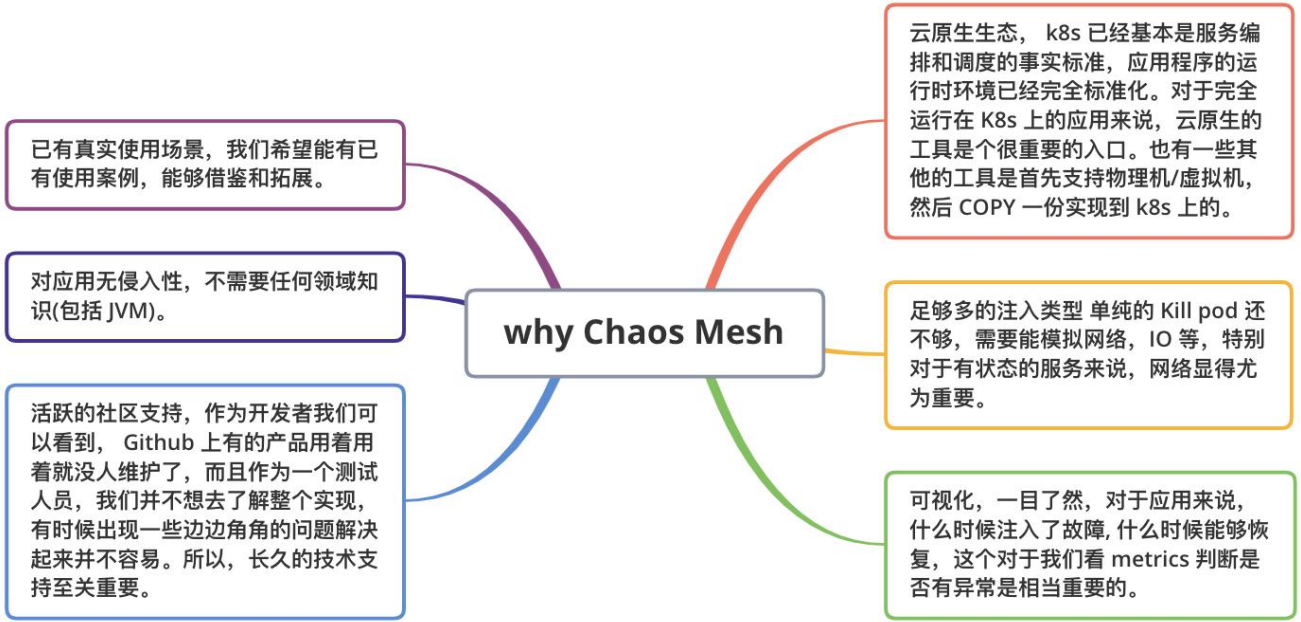
	chaos-mesh(v1.0.1)	chaosmonkey(v2.0.2)	chaosblade(v0.5.0)	chaoskube(v0.19.0)	litmus(v1.3.0)
Platform supported	K8s	VMs/ Container	JVM/Container/ K8s	K8s	K8s
CPU burn	✓	✗	✓	✗	✓
Mem burn	✓	✗	✓	✗	✓
container kill	✓	✓	✓	✗	✓
pod failure	✓	✗	✗	✗	✗
pod kill	✓	✗	✓	✓	✓
network partition	✓	✗	✗	✗	✗
network duplication	✓	✗	✓	✗	✗
network corrupt	✓	✗	✓	✗	✓
network loss	✓	✗	✓	✗	✓
network delay	✓	✗	✓	✗	✓
I/O delay	✓	✗	✓	✗	✗
I/O errno	✓	✗	✓	✗	✗
Disk fill	✓	✗	✓	✗	✓
Disk loss	✓	✗	✓	✗	✓
Time skew	✓	✗	✗	✗	✗
Kernel chaos	✓	✗	✗	✗	✗



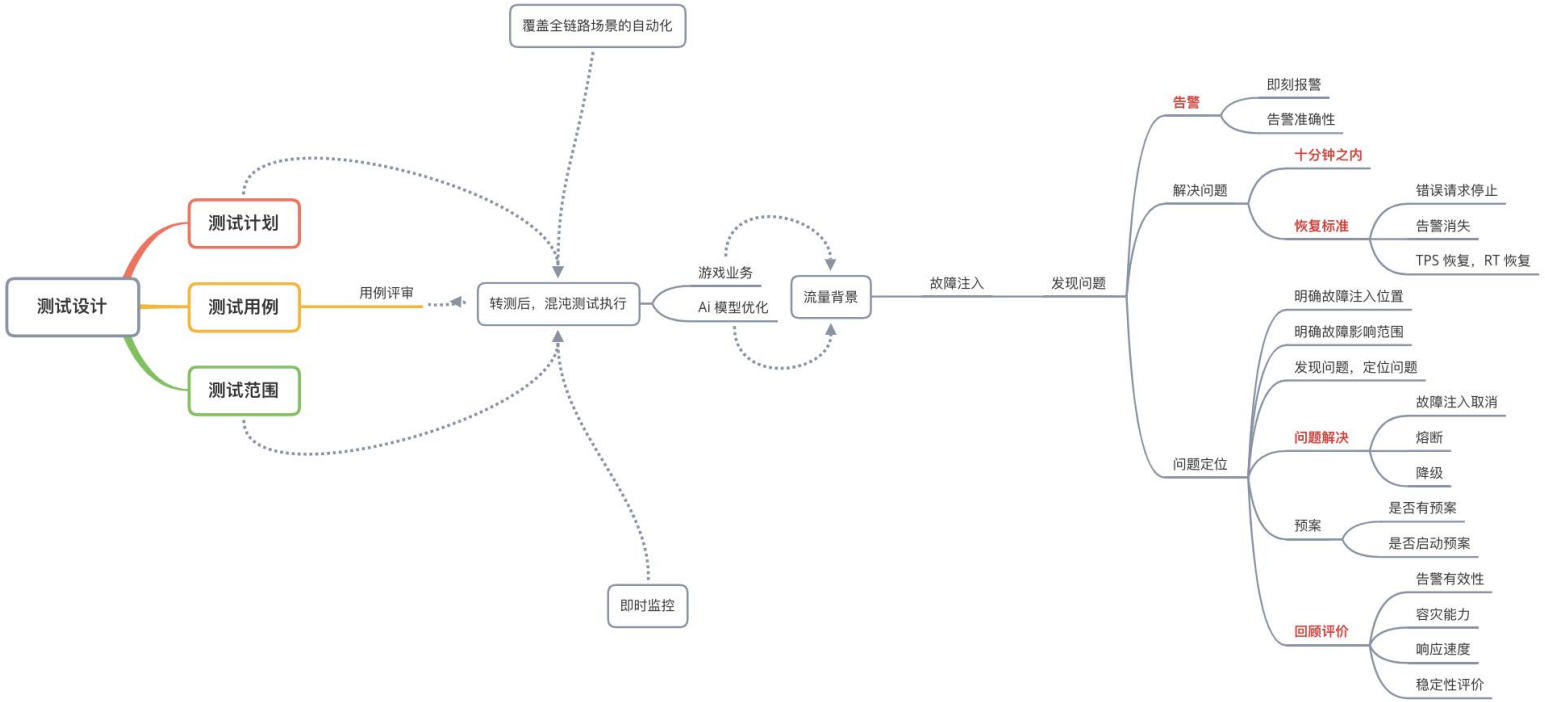
- PodChaos: kill / fail / container/...
- NetworkChaos: delay / lose / dup / partition / ...
- IOChaos: latency / fault / ...
- TimeChaos: clock skew
- KernelChaos: kernel fault injection
- StressChaos: burn cpu and memory
- DNSChaos ....

- Controller Manager
- Chaos Daemon
- Chaos Dashboard
- Grafana datasource plugin

为什么是 Chaos Mesh



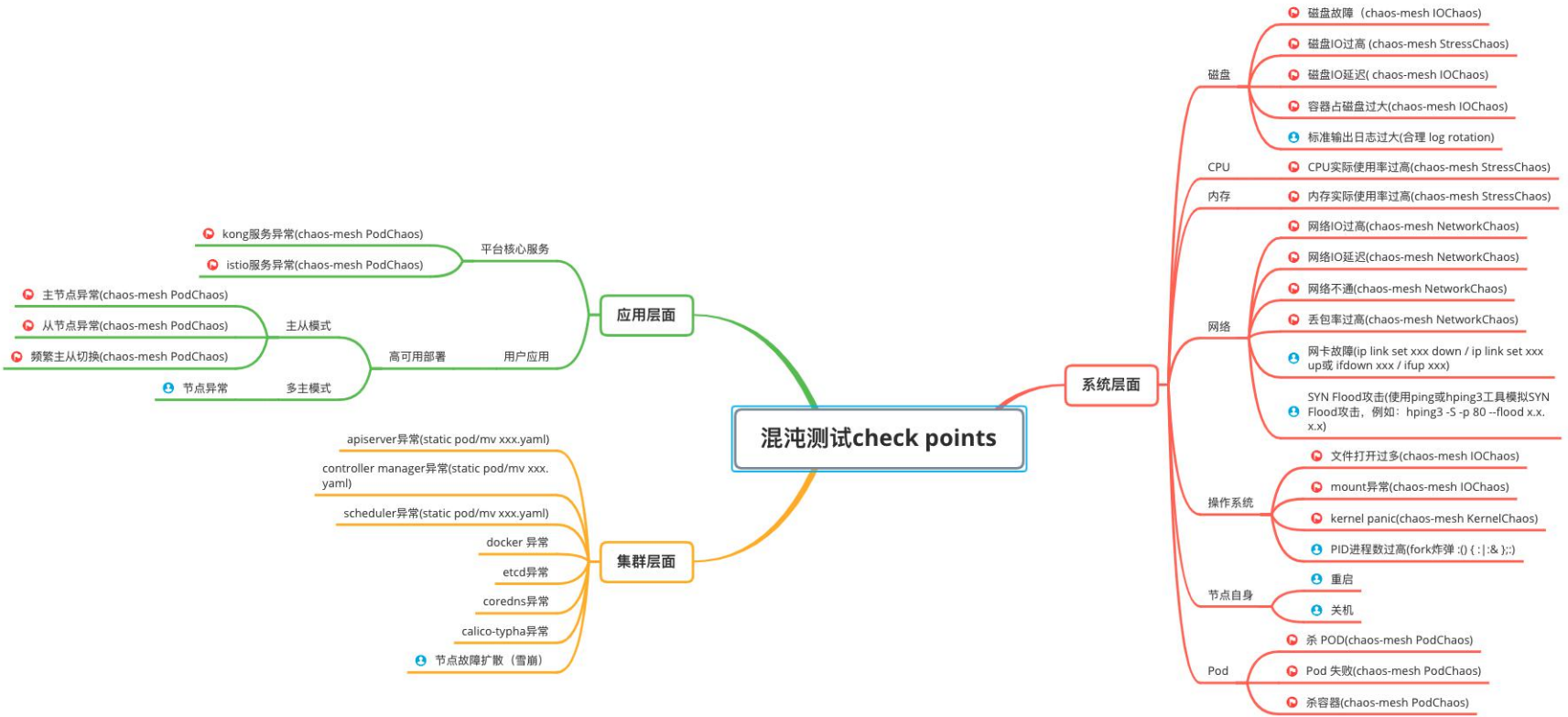
Chaos Mesh在网易伏羲的实践



Chaos Mesh在网易伏羲的实践

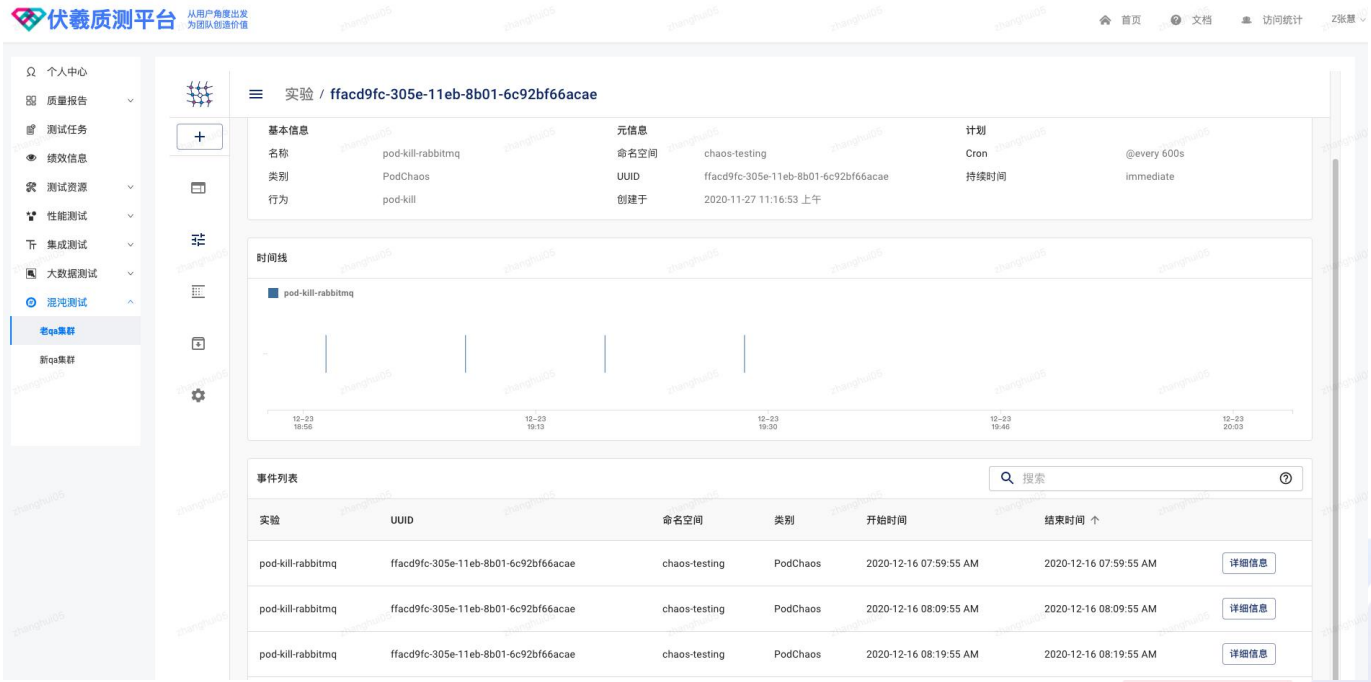
服务部署类型	故障方式	测试目标	备注
deployment	chaos-mesh	如：kong、istio、coredns	
statefulset	chaos-mesh	如：rabbitmq、redis	
pod	chaos-mesh	如：training、其他	
static pod	1、docker ps--> docker mv 2、mv xx.yaml	如：apiserver、controller manager、scheduler	qa环境目录： /etc/kubernetes/manifests
daemonset	chaos-mesh	如：calico、kube-proxy、nvidia	
物理部署	宕机、关机、重启		

Chaos Mesh在网易伏羲的实践



Chaos Mesh在网易伏羲的实践

云原生社区 Cloud Native Community




```yaml
apiVersion: chaos-mesh.org/v1alpha1
kind: PodChaos
metadata:
  name: pod-kill-rabbitmq
  namespace: chaos-testing
spec:
  action: pod-kill
  mode: one
  selector:
    namespaces:
      - danlupre
    labelSelectors:
      "app": "rabbitmq"
  scheduler:
    cron: "@every 600s"
```

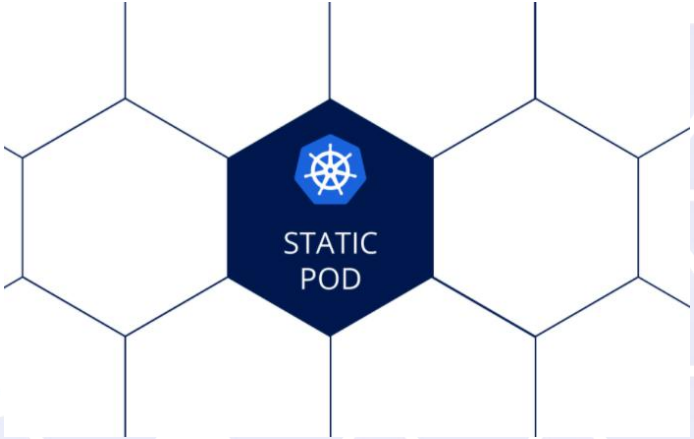
Chaos Mesh在网易伏羲的实践

云原生社区 Cloud Native Community

比如：节点异常

定时触发宕机
chmod u+x chaos-node.sh

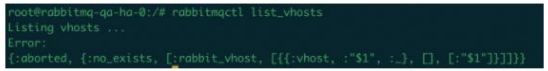
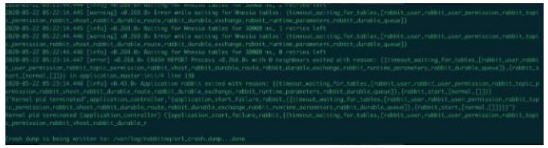
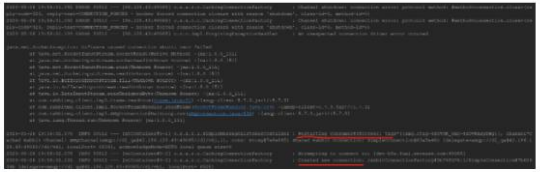
比如：static pod 异常
定时 mv statics-pod.yaml



Chaos Mesh在网易伏羲的实践

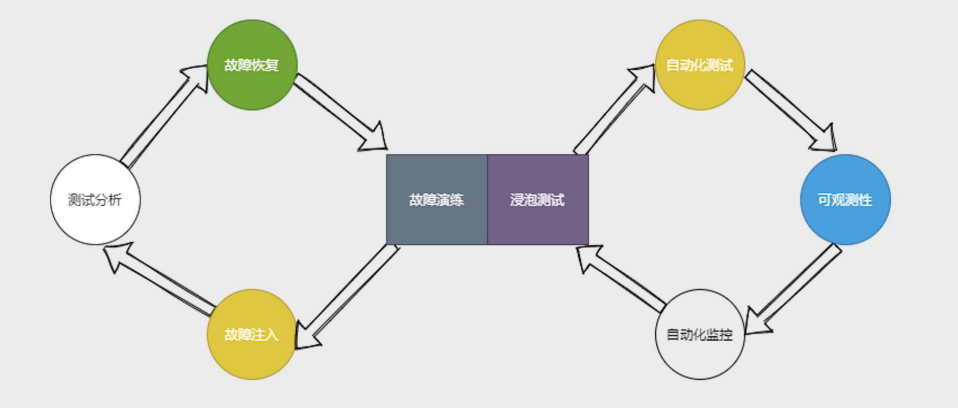
云原生社区 Cloud Native Community

提前暴露**30+**风险问题

编号	问题描述	问题原因	测试方案	解决方案
1	官方术语Cluster Network Partition，或Split-Brain 	broker节点宕掉后，net tick time(默认60s)超过后，	600s随机kill一个pod --grace-period=0 --force	参照官方文档，关于“partition handling strategies”部分，涉及三种auto handling策略。 这里考虑融入autoheal策略
2	Error: {:aborted, {:no_exists, [:rabbit_vhost, [{{:vhost, :"$1", :_}, [], [:"$1"]}]]}} 		600s随机kill一个pod --grace-period=0 --force	这种情况目前看是down掉的broker node还没起来或者上没有join到集群导致
3	启动失败 	结合mnesia源码部分(mnesia:wait_for_tables)，宕掉的rabbitmq节点在启动 app的时候，要么hang直到所有表可访，要么超时，其它细节可参见（图1）。	600s随机kill一个pod --grace-period=0 --force	这个问题，通过引入initContainer，对PV下的mnesia db进行清理操作，目前镜像yaml已更新，且运行后没有在遇到此类故障
6	broker node宕，引发connection Channel shutdown情形之一 	就目前出现的异常片段时机，初步判断是当前宕掉的broker node是客户端连接的一个endpoint	本地模拟混沌测试策略，300s随机kill一个pod，同时每隔5s，发布一条消息到broker队列	
7	broker node宕，引发connection Channel shutdown情形之二 	同上，另外，上述判断主要根据节点状态以及实时日志的实际情况	本地模拟混沌测试策略，300s随机kill一个pod，同时每隔5s，发布一条消息到broker队列	结合(图 3)，大致判断此类情形不会造成消息发送失败

Chaos Mesh在网易伏羲的实践

For FuXi , 结合流量回放在测试环境真实模拟用户流量；或真实线上集群上混沌；



For Chaos Mesh , 通用组件的开箱即用；比如 Redis、Kafka等等

欢迎对混沌测试，对私有云稳定性感兴趣的同学一起关注云原生社区 stability SIG 、Chaos Mesh 社区和网易伏羲 ~


云原生社区