

The SACC logo is rendered in a bold, white, sans-serif font with a blue glow effect. It is positioned in the upper right quadrant of the image, above the main conference title. The background features a blue wireframe architectural design with a perspective view of a city skyline and a large gear-like structure at the bottom left.

2021 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2021

数字转型 架构重塑

IT168.com

ChinaUnix

ITPUB

云上会议 网络直播 | 2021.5.20-2021.5.22

盗梦未来，浙江移动线上混沌工程的实践之路

浙江移动 史军艇

目录

浙江移动 史军艇

- 浙江移动 SRE团队
- “盗梦”演练平台 负责人
- 专注稳定性体系建设

01 浙江移动在云原生架构下的稳定性挑战

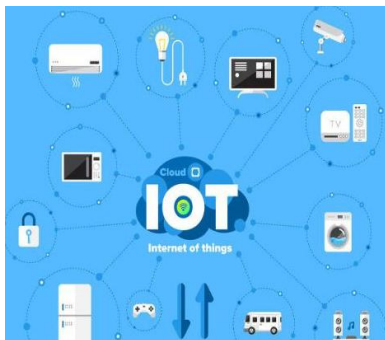
02 传统企业落地混沌工程的步骤

03 混沌平台产品化演进

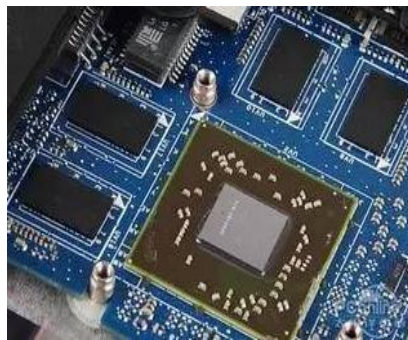
04 落地场景及案例分享

05 持续演进及思考

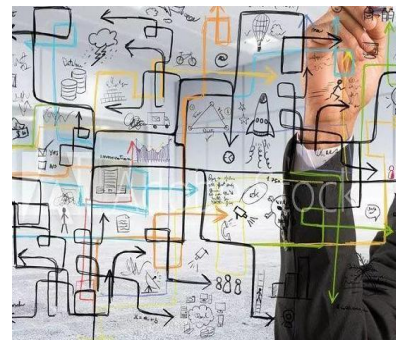
运营商系统的特点及挑战



用户规模



底层异构



业务长尾



新型需求

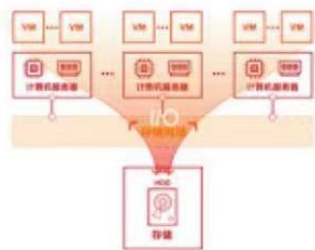
- IOT拓展, 使得用户规模指数增长
- 底层基础资源种类复杂, 国产化硬件替代
- 业务架构复杂, 存在长尾效应, 垂直烟囱的紧耦合
- 秒杀、直播等新型营销模式的需求扩充

围绕技术架构演进和数字化转型, 卸下传统企业繁重的历史包袱

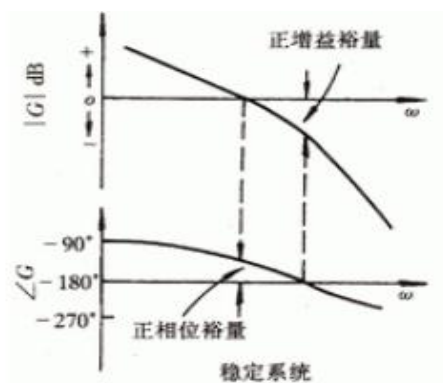
浙江移动云原生演进历程



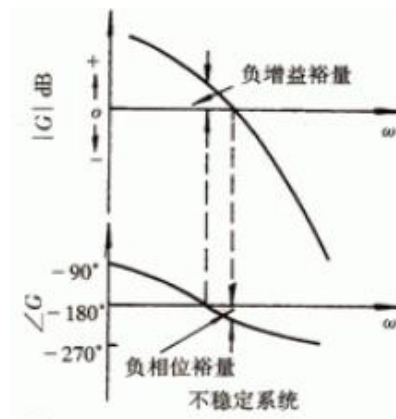
为什么要使用混沌工程



原子时代的普通系统



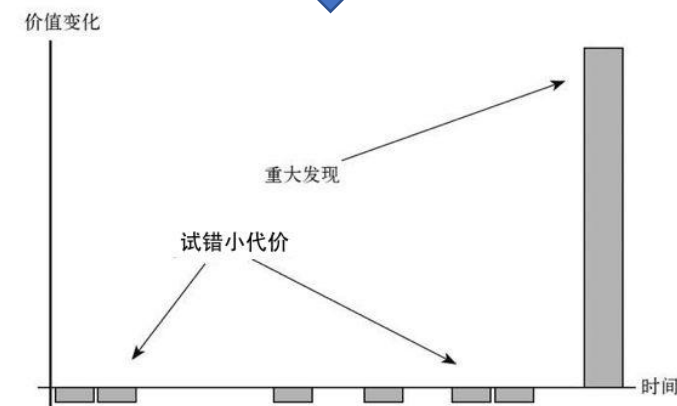
比特时代的复杂系统



云原生应用架构自变量

- 基础设施出现问题对业务的影响
- IAAS服务集群出问题对业务的影响
- PAAS服务集群出问题对业务的影响
- 微服务组件出问题对业务的影响
- 多可用区部署、泳道隔离机制
- 自动扩缩容机制
- 服务注册、依赖问题
- 运维监控平台问题

.....



没有一个架构师可以完全掌握所有的调用关系，实践是检验真理的唯一手段

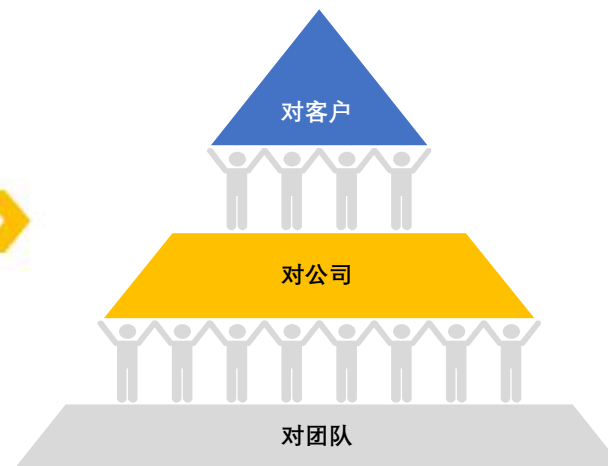
混沌工程的意义



故障驱动



“故障驱动”



提升用户感知
扩大用户基数



避免发生重大故障
抵御“黑天鹅”事件
减少收益损失
投资回报率最大化



提升组织技术能力
提升团队作战能力
打造1-5-10



图片引自《盗梦空间》

人类总是幻想能穿越时空去修复历史故障
混沌工程提供了可能！

目录

01 浙江移动在云原生架构下的稳定性挑战

02 传统企业落地混沌工程的步骤

03 混沌平台产品化演进

04 落地场景及案例分享

05 持续演进及思考

落地步骤-分层故障注入能力

逻辑节点	分类	场景
应用节点	进程实例	进程宕、进程重启
		进程HANG
		进程cpu、内存、IO异常
		业务线程池满
		连接数高
	应用代码	频繁fullgc
		特定方法类延时
		特定方法返回异常
		特定方法参数修改
		数据库切换
数据库层		数据库阻塞、大量锁表
		数据库单节点服务中断
		数据库主机单点挂掉
		数据库主机目录满
		数据库连接数满
		数据同步延时
		索引失效
		大量慢SQL
		Redis主从切换
		Redis宕机、重启
缓存中间件 (Redis、Memcached)		Redis内存耗尽
		Redis连接打满
		Redis哨兵重启
		Redis遍历Keys值
		Redis流量异常
		Redis KEY值缺失、AOF异常
		slave关闭异常
		slave网络丢包
		master网络丢包
		Redis内存配置异常
技术组件节点	Nginx	缓存击穿
		Memcached宕机
		Nginx集群重启
		Nginx反向代理异常
		Nginx日志被打满
		上述组件故障
	
		故障注入
		故障注入
		故障注入
		故障注入

故障注入



应用架构

故障注入



PAAS服务集群

故障注入



IAAS服务集群

故障注入



网络设备

故障注入



物理机/虚机

MSP控制节点	API网关	缓存读取部分数据丢失
		网关服务重启、切换、宕机
		API网关线程池打满
		网关服务响应延时
		网关服务响应延时、返回值异常
	istio网格	在线扩缩容
		网关服务流控失效
		大请求风暴
		控制节点异常、网络抖动、通信延时
		pod响应延时、返回异常
容器云控制节点		部分sidecar路由异常
		marathon部分节点中断
		mesos部分节点中断
		kubernetes控制面部分节点中断
		孤儿容器
		删容器、停容器
		杀pod、停pod
		主端防火墙宕机
		主端交换机宕机
		主端交换机单节点停服
网络设备节点	防火墙	负载均衡主端端口失效/异常
		特定IP流量不通或单通
		主端DNS解析异常
		特定链路网络抖动、丢包、延时
		虚拟机宕机、重启
	交换机	虚拟机资源耗尽(CPU、内存、IO)
		虚拟机网卡延时
		文件句柄满
		Linux时钟频率异常
	
		故障注入
		故障注入



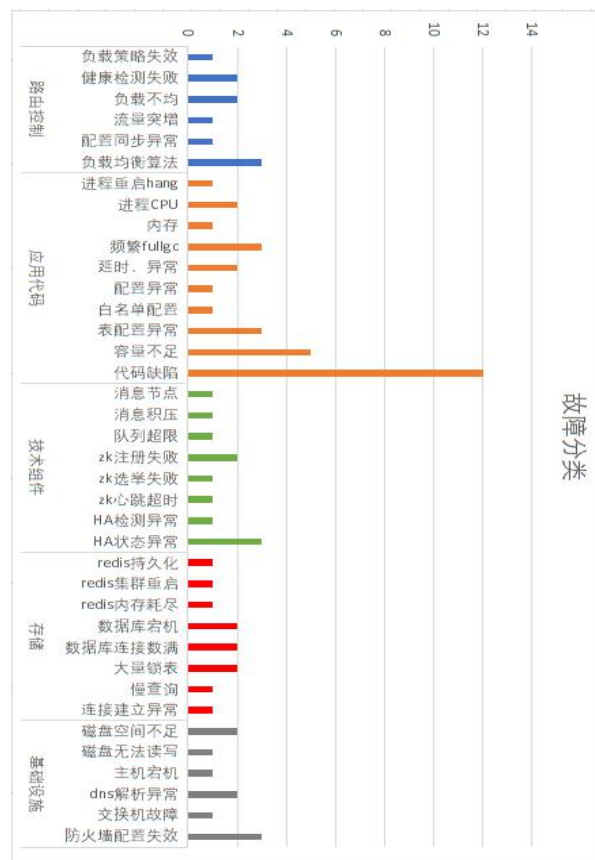
故障注入



监控节点	监控组件	故障注入
	ELK日志服务响应延时	
	kafka消息积压	
	jstom连接数过高	
	jvm信息采集延时	
	

有序的注入层次，无序的注入结果！

落地步骤-历史故障画像



故障复盘机制



Common Scenes

Route	负载策略不合理、失效	健康检测失效	负载不均	流量突增	负载配置同步异常
	进程重启、hang	进程cpu、IO、内存高、fullgc频繁	业务线程池满、容量不足	特定方法延时、异常	配置项异常
Application	消息节点不可用	消息积压	队列超限	HA检测失效	HA状态数据异常
	zk注册失败	zk选主异常	zk注册延时	zk目录写入脏数据	zk心跳超时
components	redis重启、宕机	redis持久化异常、缓存击穿	redis内存耗尽	数据库连接数满、连接建立异常	数据库锁表、慢查询
	cpu高压、打满	内存占满	IO高	能耗尽	时钟不同步
store (data)	mesos调度失效	etcd集群异常	kubelet等控制平面通信异常	pod注册失效、错乱	marathon接口异常
	网卡满	网络抖动、丢包、延时	DNS解析异常	交换机错包、故障	防火墙配置失效
computation (OS)	机房断电	机房断网	主机宕机、重启	交换机、硬负载端口失效	磁盘满、坏
infra structure					
physical device					

历史总会惊人的相似，前车之鉴是最容易获得的高价值结果

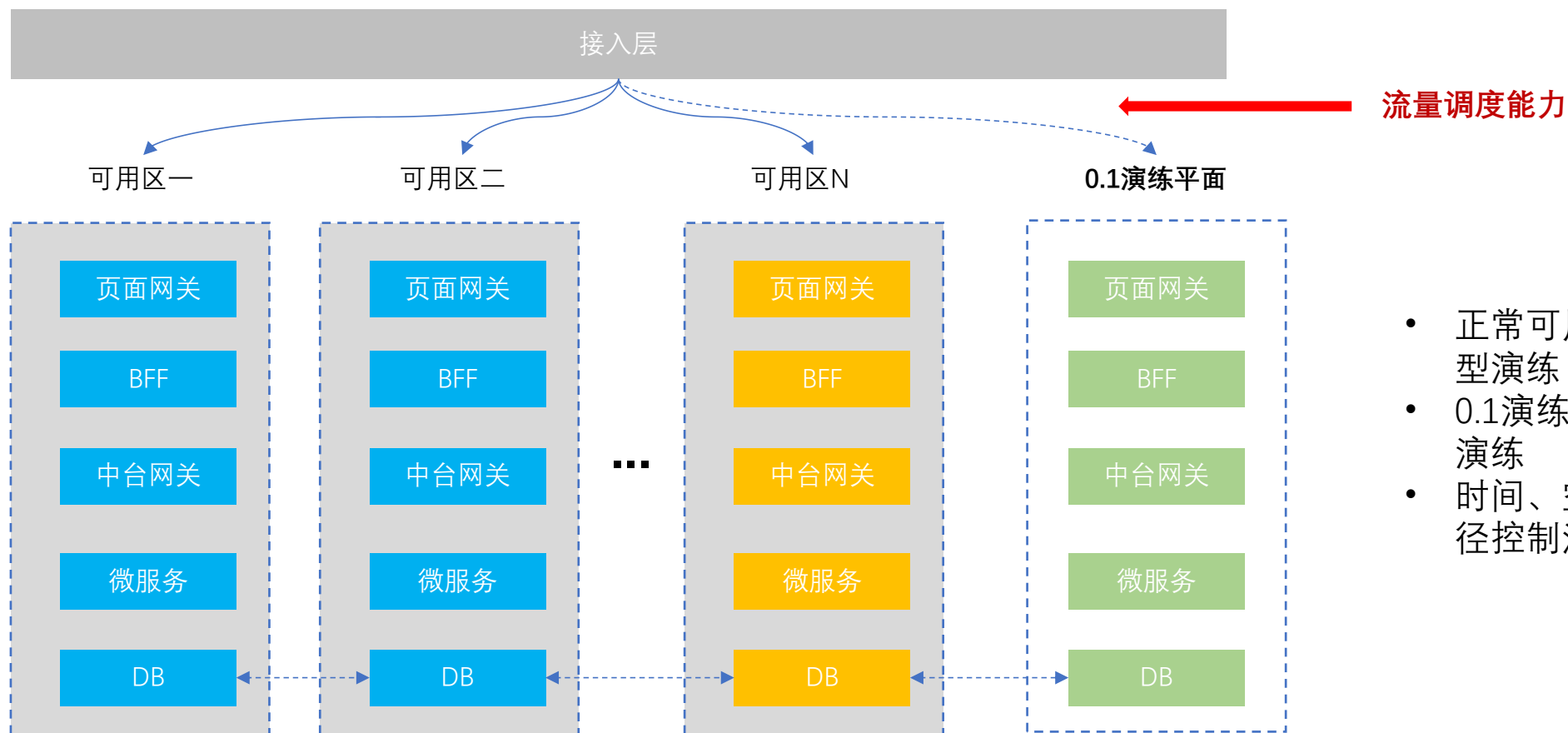
落地步骤-实验场景的系统化探索



✓ **高价值场景为导向**
以泳道逃生、任意节点可重启可宕机为出发点，用**失效影响分析 (FEMA)**自底而上，探寻所有具有**高频发生率、重大风险**的故障点。

功能点	故障模式	故障影响	严重程度	故障概率	风险程度	对抗/纸牌
登陆、受理	服务网关redis存在bigkey, 导致redis无法访问	网关接口调用失败, 导致大部分用户登陆、受理失败	高	中	高	对抗
受理	产品缓存redis磁盘空间不足, AOF模式会导致日志无法写入及缓存数据无法备份	读取产品失败, 导致大部分用户套餐受理失败	高	中	高	纸牌
受理	zookeeper写入失败	中心化服务启动失败, 小部分用户受影响	中	中	中	对抗

落地步骤-多可用区、流量调度、演练平面



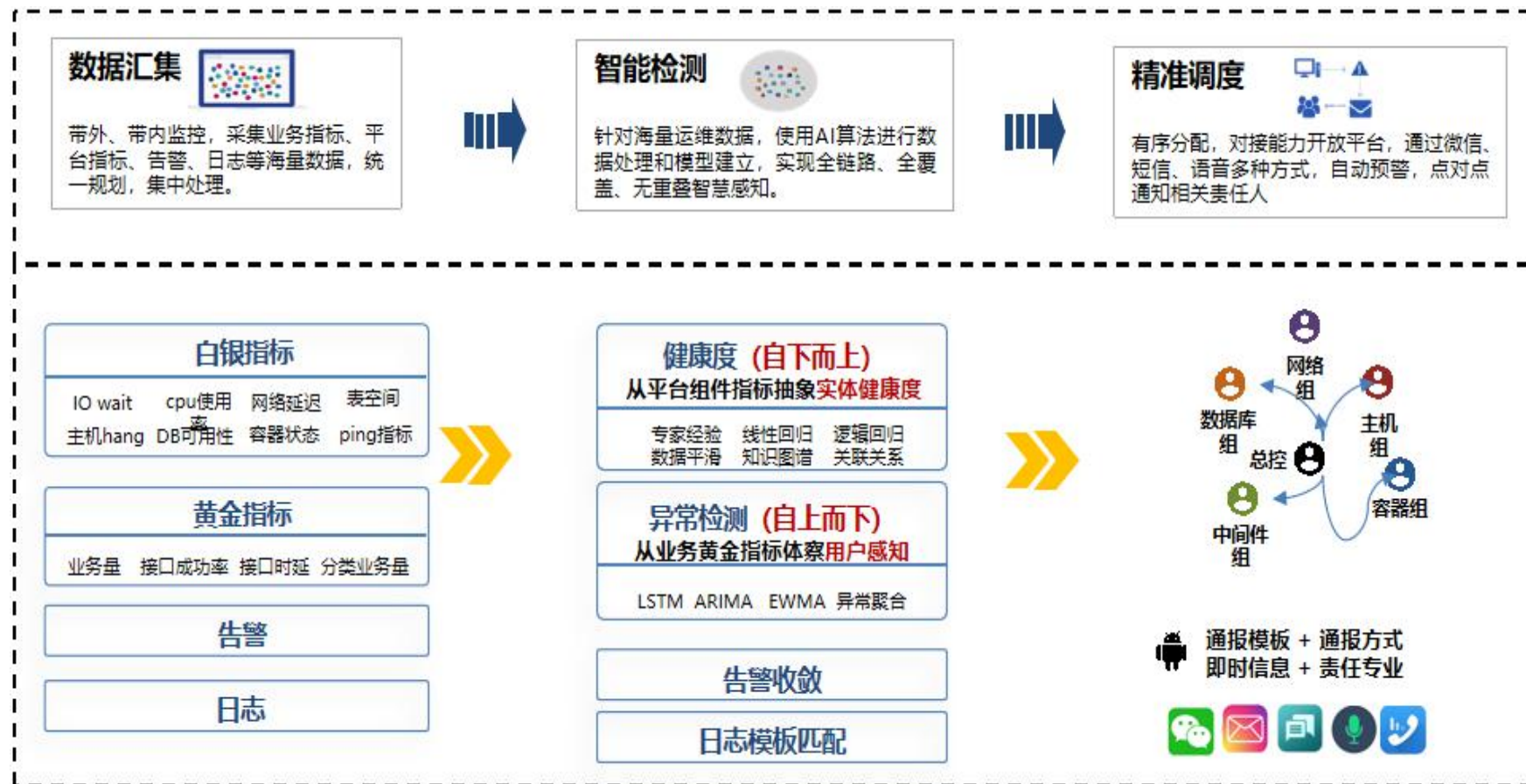
- 正常可用区进行晚上大型演练
- 0.1演练平面白天无间断演练
- 时间、空间最佳爆破半径控制法则

落地步骤-流量回放压测



- 通过流量回放压测，在任意可用区、任何时刻都可以发起生产流量，来验证故障注入的效果

落地步骤-监控矩阵、分层自愈



业务数据级
(半自动、辅助决策)

应用服务级
(应急、半自动、辅助决策)

平台网元级
(常规、自动决策)

基础设施级
(常规、自动决策)

可观测能力和自愈能力越强，发现的意外惊喜会更多！

目录

01 浙江移动在云原生架构下的稳定性挑战

02 传统企业落地混沌工程的步骤

03 混沌平台产品化演进

04 落地场景及案例分享

05 演进思路及方向

产品化演进-总体架构

业务形态

日常演练

联合演练

红蓝对抗

自动化演练

平台能力

架构感知

场景生成

流量生成

故障注入

监控护栏

采集观测

权限审批

流水线演进

准备

评估

执行

观测

清场

报告

注入框架

cmcc-chaos

chaosblade

chaos mesh

基础设施

物理机

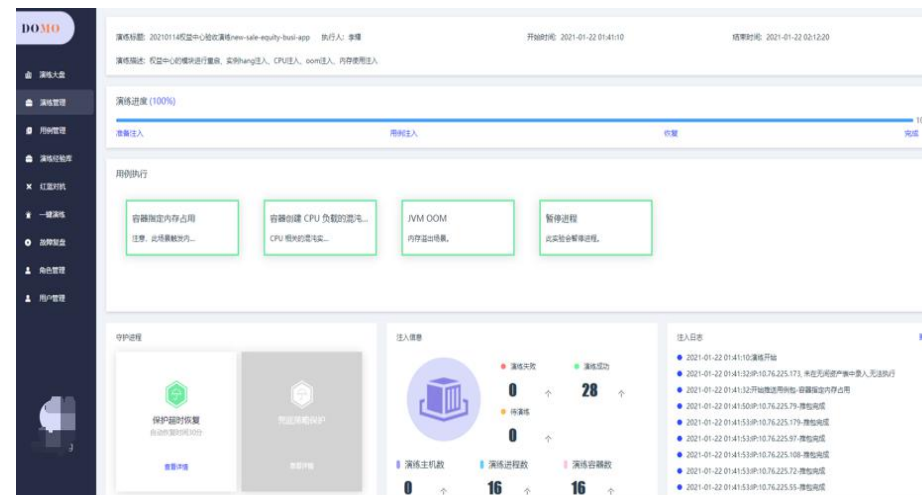
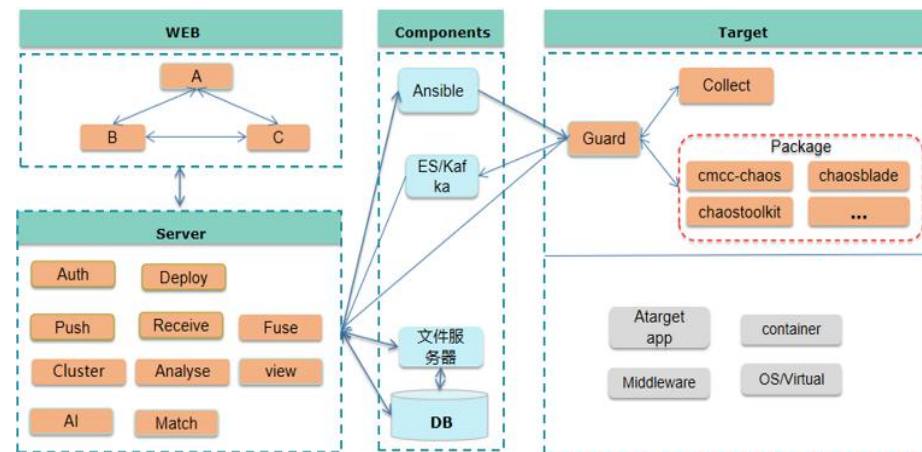
虚拟机

Mesos

Marathon

Kubernetes

IDC机房

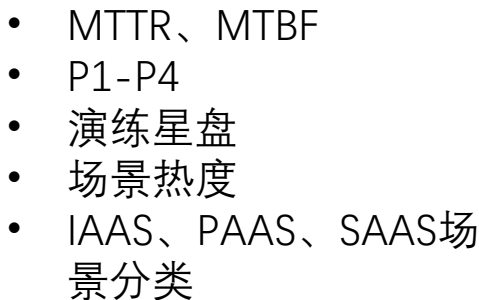


产品化演进-沉淀演练经验库

The screenshot displays a web interface for a drill experience library. On the left is a dark sidebar with navigation links: 演练大盘, 演练管理, 用例管理, 演练经验库, 红蓝对抗, 一键演练, 故障处置, 角色管理, and 用户管理. The main area shows a grid of 16 drill cards, each with a title, a brief description, and a difficulty level (基础, 进阶, 高级). The cards are organized into four columns and four rows. At the bottom, there is a pagination bar showing '1' to '8' and '下一页', along with '共160 数据' and '确定'.

卡片标题	描述摘要	难度等级	主题分类
Redis实例节点内存耗尽	对redis集群中的某一个节点进行OOM-KILL, 使得该节点内存耗尽, 导致业务产生异常	基础	技术组件故障主题
容器CPU暴增	选择指定的容器, 对其进行CPU暴增注入	基础	微服务故障主题
日志线程和业务线程异步	日志线程和业务线程异步, 然后采用线程池的方式, 我们采用日志线程异步的方式, 然后采用线程池的方式, 我们采用日志线程异步的方式	基础	应用故障故障主题
zk集群主从切换	模拟zk集群主从切换, 导致zk集群主从切换失败, 导致业务异常	基础	技术组件故障主题
进程方法超时+isee平台调用链监控不及时	模拟场景, 在业务工具出现异常的情况下, 出现方法链的超时, 方法链进行超时操作, 并增加isee平台的调用链监控, 导致数据失真	进阶	应用故障故障主题, 监控故障主题
分布式调度平台产生孤儿容器	涉及分布式的调度平台, 比如业务容器调度过程中的某一个节点的调度, 涉及到分布式调度平台, 比如业务容器调度过程中的某一个节点的调度	进阶	容器云组件主题
yaml文件在线修改	对yaml文件的yaml文件进行修改, 修改不正确的配置项, 可能会导致不正确的业务逻辑	进阶	service-mesh主题
ZK数据存储服务写入数据慢	模拟zk集群的数据写入慢, 导致写入数据失败, 导致业务异常	进阶	技术组件故障主题
HDFS节点存储资源耗尽	向HDFS集群的主机中存储资源中注入大量数据, 导致存储资源耗尽, 导致业务异常	基础	技术组件故障主题
主DNS解析成功率骤降	主DNS解析成功率骤降, 导致业务异常, 导致业务异常	基础	智能网络故障主题
API网关服务返回异常	API网关服务返回异常, 导致业务异常, 导致业务异常	基础	应用故障故障主题
大量非法Token请求 (针对API网关)	针对API网关进行非法token请求, 导致网关服务异常, 导致业务异常	进阶	MSP技术中台主题
java应用发生OOM异常	通过调用JVM内存空间, 导致JVM内存空间耗尽, 导致业务异常	进阶	微服务故障主题
网络负载掉电后恢复	网络负载掉电后恢复, 导致业务异常, 导致业务异常	进阶	智能网络故障主题
MQ时序性消息积压验证	MQ时序性消息积压验证, 导致业务异常, 导致业务异常	进阶	技术组件故障主题
业务线程阻塞	业务线程阻塞, 导致业务异常, 导致业务异常	进阶	微服务故障主题
数据库主备同步延迟	数据库主备同步延迟, 导致业务异常, 导致业务异常	基础	基础故障
istio控制节点入口Ingress-gateway异常	istio控制节点入口Ingress-gateway异常, 导致业务异常, 导致业务异常	基础	应用故障故障主题
DB部分实例连接打满	模拟数据库实例连接打满, 导致业务异常, 导致业务异常	基础	基础故障
容器内存使用进程返回异常	通过容器内存使用进程返回异常, 导致业务异常, 导致业务异常	进阶	应用故障故障主题

- 基础-进阶-高级的演练体系
- 8大场景主题



目录

- 01 浙江移动在云原生架构下的稳定性挑战
- 02 传统企业落地混沌工程的步骤
- 03 混沌平台产品化演进
- 04 落地场景及案例分享
- 05 演进思路及方向

案例一：分布式任务调度平台孤儿容器实验

实验目的：

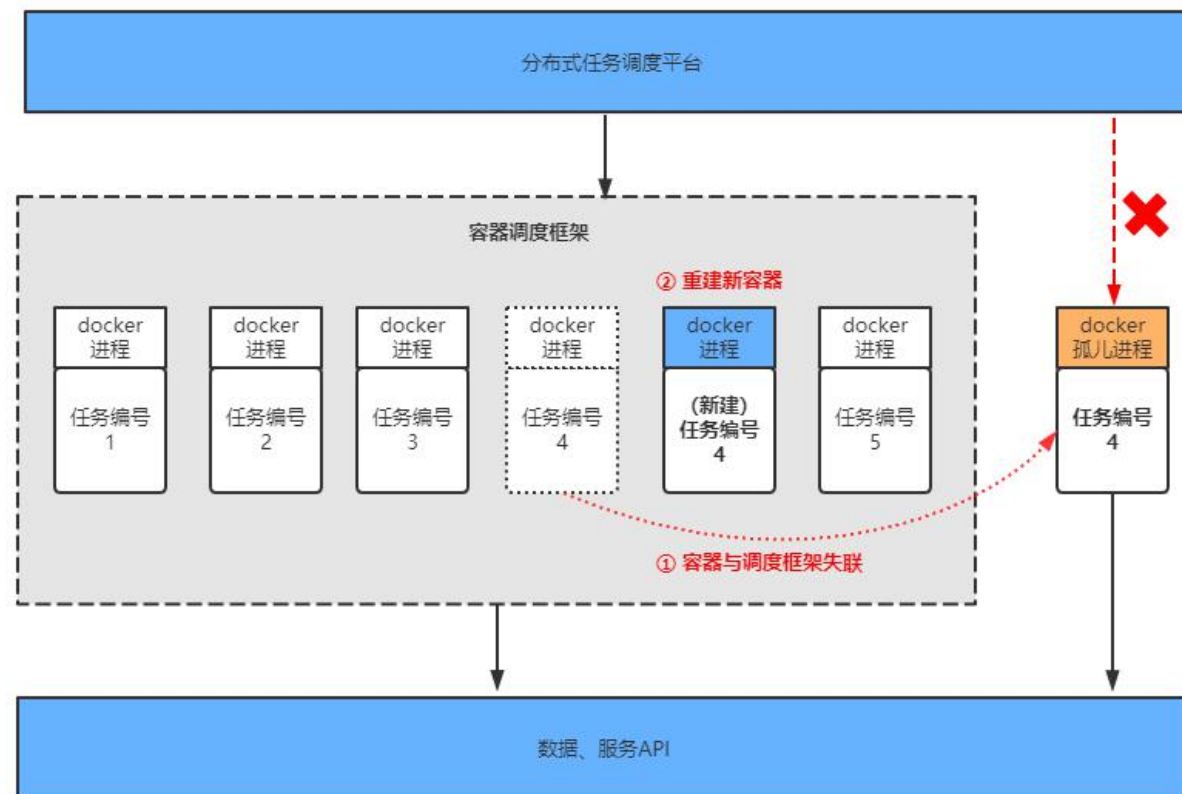
验证在容器云平台调度组件出异常时，容器云平台本身以及ET分布式调度框架是否具备重复任务的容错机制，保证业务的准确性。

实验结果：

containerd进程出现资源异常重启之后，无法调度到历史的容器id，这样就导致分布式调度平台去销毁某个任务容器时，调用marathon、mesos接口成功，但是最终调用dockerAPI失败，导致老的容器销毁失败，结果是ET平台又重新拉起一个相同任务的容器产生批量重复数据。

Active Tasks

Framework ID	Task ID	Task Name	Role	State	Health	Started	Host
...690c7b2-0009	ioecrrmtoesop572.741ee348-c63c-11ea-91bc-0242b7	rmtoesop572	*	RUNNING	-	7 hours ago	Sandbox
...690c7b2-0011	ioecrrmtoesop572.437bf87e-c31c-11ea-af9e-02426e	rmtoesop572	*	RUNNING	-	4 days ago	Sandbox



实验成果：

- (1) **开源组件BUG**：容器云组件版本开源bug导致containerd子进程异常重启的场景下无法识别之前拉起的容器，目前已替换所有bug版本。
- (2) **容器平台查杀能力**：通过自研改进来增强平台的抵抗能力，在孤儿容器产生的第一时间进行自毁查杀。
- (3) **应用平台设计规范**：在应用研发流程中，完善分布式调度平台增加应用心跳注册的机制。

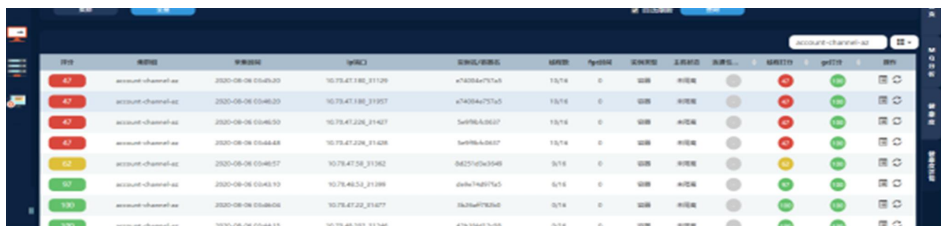
案例二：多服务中心耦合性的雪崩实验

实验目的：

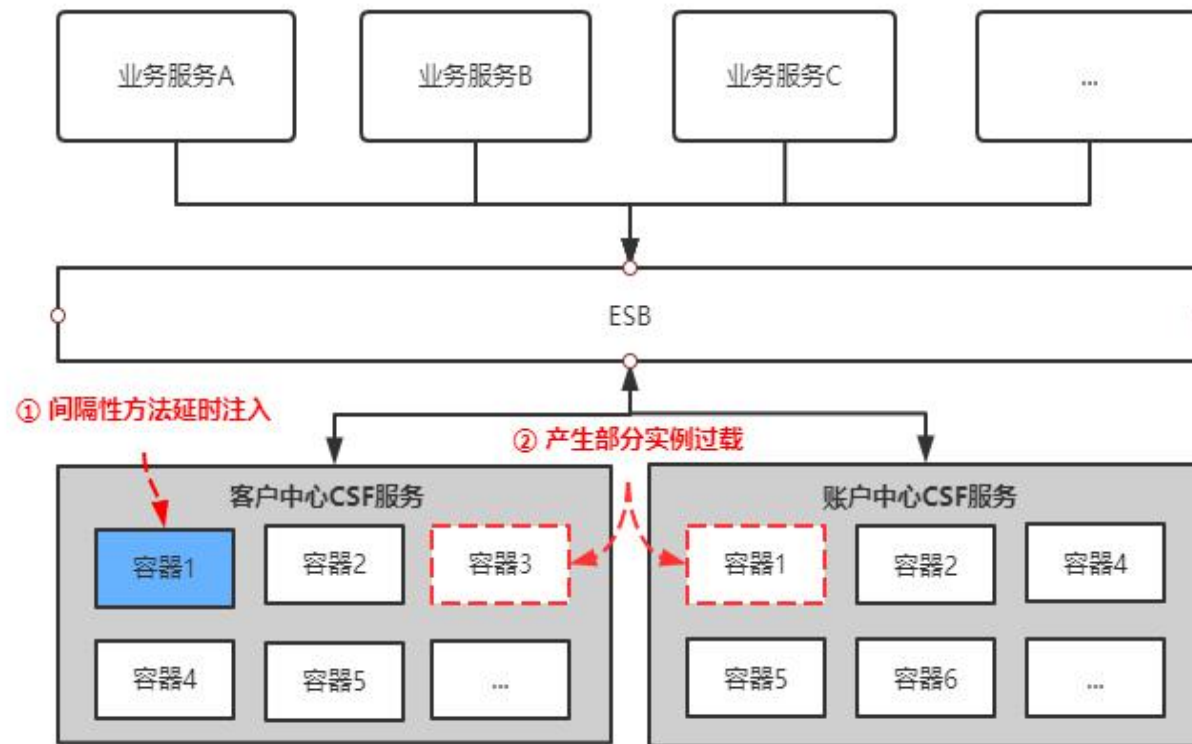
应用链路的设计是否满足高内聚低耦合原则，通过中心化服务个别节点异常后可能引发的雪崩效应验证，另，自愈模型是否可以把雪崩场景控制在业务无感知的范围内。

实验结果：

个别实例的性能低下，最终导致了5个实例左右的实例非健康，即在雪崩效应的萌芽期，通过无间平台自愈恢复。



实例ID	实例名称	实例地址	实例IP	实例端口	实例状态	实例健康	实例心跳	实例存活	实例存活时间
47	account-channel-ec	2020-08-08 05:46:20	10.79.47.182_31126	470004717165	10/14	0	健康	未健康	10/14
47	account-channel-ec	2020-08-08 05:46:20	10.79.47.182_31127	470004717165	10/14	0	健康	未健康	10/14
47	account-channel-ec	2020-08-08 05:46:50	10.79.47.228_311407	470004717165	10/14	0	健康	未健康	10/14
47	account-channel-ec	2020-08-08 05:46:48	10.79.47.228_311408	470004717165	10/14	0	健康	未健康	10/14
47	account-channel-ec	2020-08-08 05:46:57	10.79.47.228_311409	470004717165	10/14	0	健康	未健康	10/14
50	account-channel-ec	2020-08-08 05:46:53	10.79.46.52_311386	470004717165	5/14	0	健康	未健康	10/14
50	account-channel-ec	2020-08-08 05:46:56	10.79.47.22_311407	470004717165	5/14	0	健康	未健康	10/14
50	account-channel-ec	2020-08-08 05:46:55	10.79.46.202_311405	470004717165	5/14	0	健康	未健康	10/14



实验成果：

- (1) 研发链路优化：通过分析业务调用链，驱动研发进行更高效的调用改造，以此来提升服务性能。
- (2) SRE自愈模型优化：改进运维平台的自愈模型，提升实例的健康度精准诊断，使得可以应对生产中绝大多数性能恶化场景。

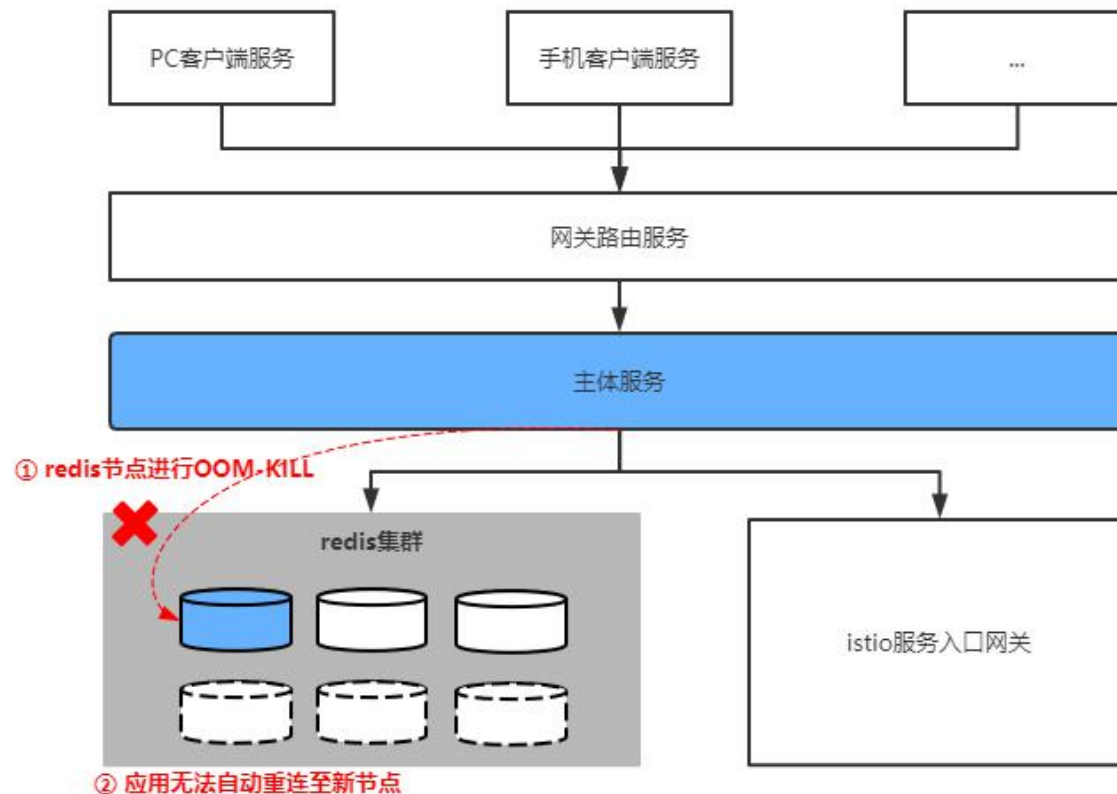
案例三：redis集群节点内存耗尽实验

实验目的：

redis组件集群部分节点内存溢出产生异常后，观察是否会影响业务的正常使用。

实验结果：

redis集群其中一个主节点进行OOM-KILL后，网关的主体服务无法有效地自动切换到所在的备用节点，最后通过快速重启故障redis节点进行恢复，导致影响业务5分钟。经查，由于主体服务中redis客户端代码bug，屏蔽了自动识别redis节点状态的配置引起。



实验成果：

- (1) 客户端代码优化：改造redis客户端代码，使得可以自动识别redis节点的可用状态，并完成切换。
- (2) 应急团队锤炼：redis节点OOM时，完成分钟级的发现并恢复。
- (3) 逃生配置失效：红军在为屏蔽单平面redis的异常，对抗中做了可用区切换，结果意外发现array的配置未同步，使得可用区成为假平面

目录

- 01 浙江移动在云原生架构下的稳定性挑战
- 02 传统企业落地混沌工程的步骤
- 03 混沌平台产品化演进
- 04 落地场景及案例分享
- 05 演进思路及展望

- 丰富传统行业特性化的演练场景
- 完善云原生演练能力，扩充service mesh的高价值场景，以应对大规模service mesh应用实施
- 提升平台的集成能力，适配各类框架，降低接入成本，统一paas标准
- 继续扩展可视化演练能力，建立完善的数字化演练体系及架构体系
- 自动化演练能力赋能微服务设计，建立从CI/CD开始的演练流水线，
- 通过自动化演练的故障标记，结合aiops模型，做高精度故障抵御、自动决策

混沌工程

蓝军战略

盗梦未来，始终跑在真正的故障之前，融合云原生自动驾驶技术，让代码对抗代码

敏捷为中心的内驱力，为企业需求迭代、业务模式创新按下快进键

从技术价值迈向商业价值

The background is a deep blue with a complex, abstract pattern of glowing wireframe cubes and rectangular prisms. These shapes are arranged in a way that creates a sense of depth and perspective, with some appearing to recede into the distance. A bright, horizontal lens flare or light streak cuts across the center of the image, passing behind the word 'THANKS'. The overall aesthetic is futuristic and digital.

THANKS