CAICT 中国信通院

# 混沌工程实践指南

（2021 年）

中国信息通信研究院云计算与大数据研究所

2021 年 12 月

# 版权声明

# 前　言

随着各行业数字化转型的快速推进，软件系统呈现出规模愈发庞大、结构更加复杂，风险点数量增多、不可预见性增强等特征。现有软件稳定性保障措施侧重于防范可预见的风险，难以适应数字化深入发展的新要求。

为探索新发展阶段下软件系统的稳定性保障手段，国外公司率先提出混沌工程理念，按照主动攻击、提前预防的思想，消除复杂系统中难以预见的风险隐患，迅速成为业内关注的焦点。我国产业界也迅速跟进混沌工程技术发展，相关工具和平台层出不穷，开源社区加快完善。

然而，目前混沌工程总体上仍处于发展初期，理论体系尚未成熟，技术工具良莠不齐，落地实践仍集中于少数头部企业。在此背景下，中国信息通信研究院云计算与大数据研究所牵头，联合业内头部企业编写本指南，梳理混沌工程的相关背景、技术要求、实践配套措施和未来的发展趋势，希望借此推动混沌工程落地实践，帮助各行业完善软件系统稳定性保障体系。由于时间仓促，水平所限，本指南仍有不足之处，欢迎联系 wangchaolun@caict.ac.cn 交流探讨。

# 目 录

# 图 目 录



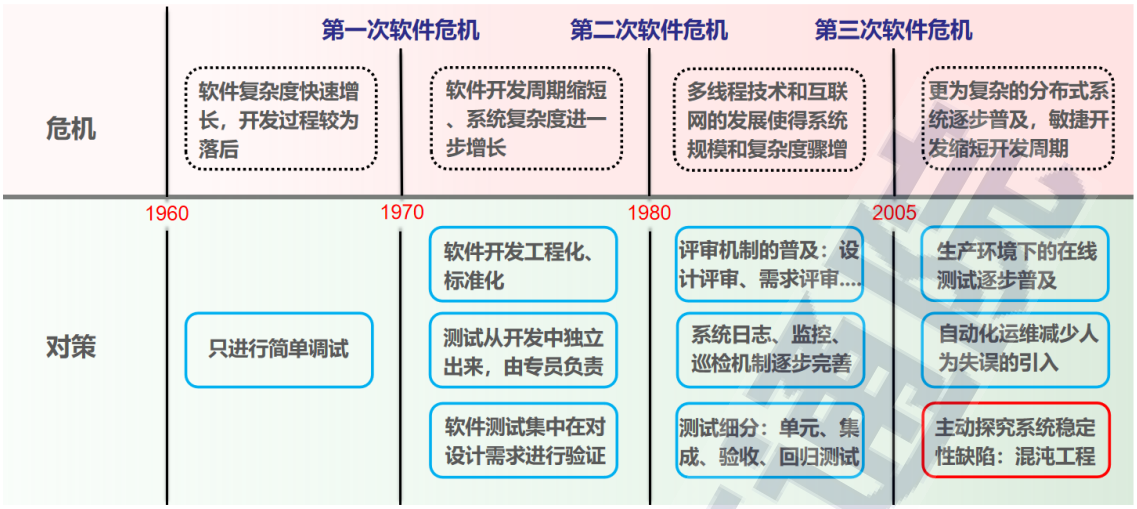
# 表 目 录

# 一、混沌工程概述

混沌工程（Chaos Engineering）是通过主动向系统中引入软件或硬件的异常状态（扰动），制造故障场景并根据系统在各种压力下的行为表现确定优化策略的一种系统稳定性保障手段。应用混沌工程可以对系统抵抗扰动并保持正常运作的能力(稳定性)进行校验和评估，提前识别未知隐患并进行修复,进而保障系统更好地抵御生产环境中的失控条件，提升整体稳定性。本章梳理了系统稳定性保障体系和混沌工程的发展历程，并给出了混沌工程实践体系框架。

## （一）混沌工程旨在主动防范软件系统的稳定性风险

自开展软件相关工作以来，从业人员和系统稳定性缺陷的斗争从未停止。随着软件系统规模扩大、复杂度增长以及开发周期缩短，历史上曾爆发多次软件危机，推动着软件从业人员不断完善系统稳定性保障措施。随着分布式系统的广泛应用以及敏捷开发、DevOps 的普及，当今软件系统在规模、复杂度和开发敏捷程度方面再次迈入一个新的阶段，系统稳定性也开始面临新的威胁，混沌工程应运而生。混沌工程作为探究系统缺陷的手段，使得软件从业人员在与系统缺陷的斗争过程中掌握主动权，很好地弥补了稳定性保障措施中的短板。




来源：中国信息通信研究院，2021 年

图 1 系统稳定性危机与对策发展时间线

## 1. 历次软件危机促使系统稳定性保障措施不断完善

1960 年起，软件系统逐步从计算机系统中分化出来，用于处理逻辑复杂的问题，其自身复杂度开始进入快速增长期，并于 70 年代爆发了第一次软件危机。软件工程理论提出后情况有所好转，但随着软件的应用范围逐步扩大，开发速度再次落后于计算机普及的速度，开发周期不断缩短、复杂度进一步增长，最终在 80 年代引发了第二次软件危机。这迫使各软件研发机构逐步建立了质量保证体系和更完善的软件工程方法。

## 2. 当前 IT 行业的高速发展带来新的稳定性隐患

随着互联网技术的发展，IT 产业在 21 世纪迎来了高速发展，于是第三次软件危机也由 2005 年开始并延续至今。在最近的十年中，软件系统在规模、复杂度和研发运维模式上均发生重大变化：

**系统分布式化后更容易受到硬件扰动的影响：**单点性能瓶颈导致

软件系统逐渐分布式化。但大型的分布式系统通常部署于成百上千个物理节点之上，其老化、损坏、连接中断概率将迅速增大，导致系统不稳定。通过增加成本购买可靠性更高的硬件设备，增强集群的监管和提高检修频率可以减少故障的发生，但却无法根除。

**现代开发体系增加了潜在缺陷引入风险概率：**一方面，现代开发体系需要多人、多部门沟通与协作，增加了缺陷引入的可能；另一方面，敏捷开发逐渐成为主流，但其在提升开发效率同时，也导致一些需要较长时间或特定外在条件才能触发的问题难以发现。

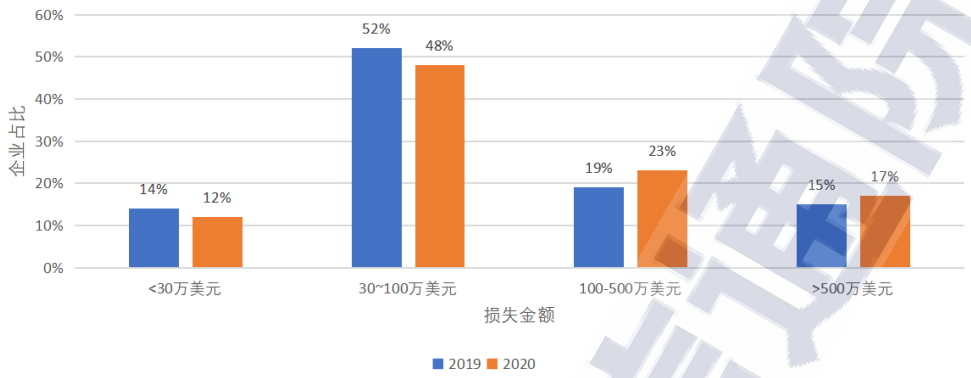
随着软件系统成为各行业基础设施，系统稳定性缺陷的破坏性也逐渐增大，并有逐年上升的趋势。近一年来便有多个大型互联网平台发生稳定性故障。行业分析机构 Statista 对世界范围内 IT 行业头部企业每宕机一小时所造成的损失进行了统计。在 2020 年，高达 40%的 IT 企业每宕机一小时的损失超过 100 万美元，比 2019 年上升 6%。其中有 17%的企业宕机一小时造成的损失超过 500 万美元。

表 1 2020 年 9 月至 2021 年 9 月影响严重的系统失效事故汇总

| 机构名称 | 发生时间 | 持续时长 | 影响范围 | 原因 |
|---|---|---|---|---|
| 哔哩哔哩 | 2021 年 7 月 | 约 1 小时 | 哔哩哔哩视频播放、直播等多项服务 | 机房故障，灾备系统失效 |
| Fastly | 2021 年 6 月 | 约 1 小时 | 包括亚马逊、纽约时报、CNN 在内的登录网页 | 系统漏洞被配置更改操作触发 |
| 推特 | 2021 年 3 月 | 约 2 小时 | 登录失败 | 系统内部错误 |
| 滴滴打车 | 2021 年 2 月 | 约 1 小时 | 滴滴打车 APP | 系统内部错误 |
| 美联储 | 2021 年 2 月 | 约 4 小时 | 美联储大部分业务 | 操作失误 |
| 谷歌 | 2020 年 12 月 | 约 1 小时 | 谷歌旗下大部分业务 | 存储超出限额 |
| 亚马逊 | 2020 年 11 月 | 约 5 小时 | 部分服务无法访问 | 系统漏洞被不当的运维操作触发 |
| 微软 | 2020 年 9 月 | 约 5 小时 | Microsoft Office 365 | 流量激增导致服 |

| | | | | |
|---|---|---|---|---|
| | | | 办公软件和 Azure 云产品 | 务中断 |

来源：中国信息通信研究院，2021 年




数据来源：Statista，2020 年

图 2 企业服务中断每小时造成的损失统计

## 3. 混沌工程是系统稳定性保障方式的新探索

现有的稳定性保障措施侧重点在于如何防范已知范围内系统缺陷的引入，对于需要特定外界扰动才能触发的故障缺乏识别和修复的手段，只能在系统故障发生时对故障进行被动的响应，导致故障应对的进度和成本不可控。

混沌工程通过主动向系统中注入可能引发故障的扰动（即软件或硬件方面的异常状态）来探究系统对扰动的承受能力，很好地弥补了稳定性保障措施中的短板。

表 2 混沌工程和现阶段稳定性保障措施的对比

| 对比维度 | 现阶段稳定性保障措施 | 混沌工程 |
|---|---|---|
| 工作内容 | 防范缺陷的引入，故障发生时对缺陷进行快速的识别和响应 | 通过实验主动探究系统缺陷 |
| 排查缺陷的类型 | 低层次缺陷，比较明显的缺陷，或已经引发故障的缺陷 | 未知的、潜在的缺陷，还未造成明显后果的缺陷 |

| | | |
|---|---|---|
| 应对缺陷的方式 | 被动响应，缺陷应对的开始时间取决于故障何时发生，缺陷应对成本不可控 | 主动响应，缺陷应对的开始时间取决于混沌工程实验时间，缺陷应对成本可控 |
| 识别缺陷的效率 | 效率低，对于一些触发条件苛刻的潜在缺陷可能需要很长时间才能被识别 | 效率高，可以使潜在缺陷尽快暴露。缩短缺陷识别周期 |

来源：中国信息通信研究院，2021 年

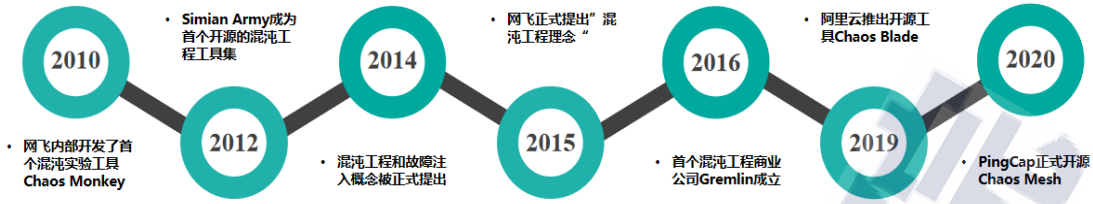
混沌工程有助于帮助系统研发人员发现系统中潜在的脆弱环节，降低稳定性缺陷可能造成的损失，提高应用上线的信心。对于不直接参与研发的团队，通过混沌工程也可以创造故障应对的场景，从而锻炼运维支撑团队发现定位故障并恢复系统的能力，找出并改进机构在稳定性保障体系中的不足。

表 3 混沌工程对系统研发运维团队不同人员的意义

| 人员类别 | 实践混沌工程的意义 |
|---|---|
| 研发工程师、架构师 | 加深对系统的理解，验证系统架构的容错能力 |
| 运维工程师 | 提高故障的应急效率，实现故障告警、定位、恢复的有效应对 |
| 测试工程师 | 弥补传统测试方法留下的空白，更主动的方式探究系统问题 |
| 产品设计人员 | 了解产品在突发情况下的表现，提升客户在突发情况下的产品使用体验 |

来源：中国信息通信研究院，2021 年

## （二）混沌工程的发展历程：国外先行、国内繁荣




来源：中国信息通信研究院，2021 年

图 3 混沌工程发展时间线

混沌工程的概念最早由网飞公司（Netflix）[1]提出。2008 年 8 月，网飞公司主要数据库发生故障，导致了长达三天的停机，造成巨大经济损失。于是网飞公司开始尝试利用混沌工程优化稳定性保障体系。其在 2010 年开发了 Chaos Monkey 程序，该程序的主要功能是随机终止在生产环境中运行的虚拟机实例和容器，模拟系统基础设施遭到破坏的场景，使得工程师能够观察服务是否健壮、有弹性，能否容忍计划外的故障。Chaos Monkey 于 2012 年在 Simain Army 项目中开源，为混沌工程工具的发展打下了基础。网飞公司在 2015 年发布了《混沌工程理念》（Principal of Chaos Engineering），主要介绍了混沌工程实验的目的、意义和方法论。2016 年混沌工程商业公司 Gremlin 成立，混沌工程正式走向商用化。

自 2018 年起，国内诸多企业也开始引入并实践混沌工程，由国内厂商主导的混沌工程开源项目 Chaos Blade 和 Chaos Mesh 在 2019 年和 2020 年被先后推出，现已发展成为具备国际顶级影响力的混沌工程项目。国内混沌工程产业繁荣发展有其必然因素：一是我国的大

1 网飞公司的主要业务包括互联网流媒体播放和 DVD 在线租赁，是世界最大的在线影片租赁服务商，用户数超过 2 亿。

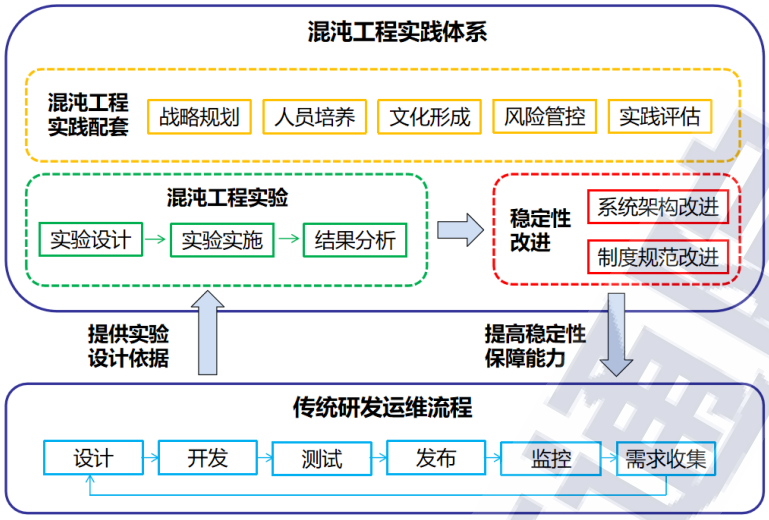
型互联网公司面对的是世界上最庞大的互联网流量，分布式软件架构、DevOps 开发模式早已大范围应用推广，分布式系统稳定性研究十分必要；二是我国拥有庞大的软件开发者群体，这使得国内混沌工程项目的开源社区较为活跃；三是我国正处于数字化转型的关键时期，对信息基础设施建设的投入巨大并有着政策支持，作为提升信息基础设施质量的手段，混沌工程也必然受到更多的关注。

### （三）混沌工程实践是系统性工作，亟需建立方法论

混沌工程并非是对传统研发运维体系的替代与革新，而是对其部分环节、部分理念的升级与延伸。混沌工程的实践落地也绝非简单的扩展混沌工程实验的规模与形态，而是一整套系统性工作。

然而由于混沌工程发展时间较短、尚未形成较为成熟方法论，国内大部分企业对混沌工程相关知识仍缺乏体系性了解，对于如何将混沌工程相关工作与现有研发运维体系相结合仍处于摸索阶段，对于如何选型支撑混沌工程实践落地的技术工具仍缺乏标准，亟需理论引导。

为助力混沌工程在我国推广落地，本指南汇总并梳理了领域内头部机构的混沌工程实践经验，并首次提出了混沌工程实践体系，共包含“一个核心工作、五项配套工作、两类延伸工作”，如图 4 所示。本小节首先对混沌工程实践体系中的各个环节进行总括性介绍，后续各章节将针对各环节内容进行详细叙述。




来源：中国信息通信研究院，2021 年

图 4 混沌工程实践体系以及其和传统研发运维的联系

### 1. 混沌工程实验是实践混沌工程的核心工作

混沌工程以实验为最小单元探究系统稳定性缺陷，因而进行混沌工程实验是实践混沌工程的核心内容。混沌工程实验设计拥有较高的自由度，通常会参考研发运维过程中遇到的问题。混沌工程实验流程有着相对固定的模式：实验开始时通过系统指令或对硬件设备进行人为干预来注入扰动；实验进行的过程中，观察并收集各指标的实时变化；实验结束后通过对比变量组收集的指标和稳态指标，评估系统的状态变化。

### 2. 战略、人员、文化、风险防范、评估体系共同构成混沌工程实践的配套体系

在组织机构中实践混沌工程需要多部门的协同配合，这就要求对混沌工程实践相关团队进行有针对性的统筹管理，建立一整套适用的配套措施，包括混沌工程相关的战略规划、人员培养、文化形成、风

险应对以及评估体系的建立。

3. 对架构和制度的两类改进方案是混沌工程实践的延伸

系统运营者在通过混沌工程发现稳定性缺陷时，对于一些架构上的隐患可以针对性地采用一些韧性设计，对于编码错误、流程失误等低层次的技术缺陷，则可在研发运维过程的制度纪律上进行改进。

## 二、一个核心工作：混沌工程实验

混沌工程实验是探究系统稳定性缺陷的最小单元，是实践混沌工程的核心要素。本章介绍了混沌工程实验在设计、实施和结果分析过程中需要关注的要点和推荐采用的步骤。

### （一）混沌工程实验设计

一个完整的混沌工程实验设计流程由建立假设、实验场景设计、系统评估指标设计、扰动类型设计、扰动注入模式设计和实验结果预期组成。

1. 建立假设

假设的建立是实验进行的基础，一切实验都是对假设进行证明或证伪。在混沌工程实验的场景下，假设通常反映的是系统可能存在的稳定性缺陷。这些假设的缺陷可以依据用户反馈、测试记录和对系统架构的理解来确立。假设的缺陷制定完成后，应逐个分析其出现概率、影响范围和严重程度来确立混沌工程实验的优先级。曾出现过并造成经济损失的假设缺陷通常是实验优先级最高的，这些有记录的故障很

有可能会再次出现，而且也可能会在结构相似的链路上引发同类型的故障。

### 2. 实验场景设计

如需观测注入扰动对系统的影响，我们需要被测系统处于一个业务场景中正常工作的状态。实验场景的设计要避免过于简单，需包含多样的任务以确保足够的覆盖范围。如条件允许，可考虑使用生产环境作为实验用场景。为了降低生产环境实验的风险，通常的预防策略是采用金丝雀（Canary）版本（发布给少量用户的试用版本）或采用流量分支作为混沌工程实验场景，以确保最坏的情况下只有一小部分用户会受到影响。如在模拟或测试环境中进行实验，实验的场景需尽量接近实际业务中的真实场景，覆盖的用户操作应尽量全面。较为有效的方法是录制生产环境中的各种变量，如流量、服务请求频率等，然后在测试中重放，或用生产环境的模拟数据进行实验场景搭建。

### 3. 系统评估指标设计

实验设计时应确立实验过程中需收集的指标，以便评估注入扰动对系统造成的影响。这些指标可根据具体的实验对象、业务场景以及可用的监控手段确立，并尽量全面，需要在系统功能或性能受损时产生明显的变化。指标确立和收集时推荐采用的做法是:

a. 关注指标平均值的同时对最优值和最差值进行收集。

b. 关注最终指标的同时对子任务指标或支撑指标进行收集。

c. 将一部分指标定为止损指标，当止损指标超出一定阈值则需终止实验，避免造成较大的损失。

典型的系统评估指标可以分为以下类别:

表 4 混沌工程实验系统指标类别

| 指标类别 | 指标描述 | 案例 |
| --- | --- | --- |
| 时间类指标 | 系统完成实验场景单个或批量任务所需的时间 | 服务器端响应时间、网络响应时间、客户端响应时间，任务完成耗时等 |
| 效率类指标 | 系统在实验场景中的工作效率 | 吞吐量，TPS（Transaction Per Second，每秒钟完成的业务数）、QPS（Query Per Second，每秒钟完成的查询数）等 |
| 失效率类指标 | 系统执行功能失败的比例 | 接口响应失败率、服务自动隔离或下线时间占比等 |
| 资源类指标 | 系统使用资源的情况 | CPU 使用率、内存使用量，磁盘输入和输出量，网络输入和输出量等 |
| 综合业务类指标 | 用户对于业务的反馈情况 | 用户重试率、用户报错数量等 |

来源：中国信息通信研究院，2021 年

## 4. 扰动类型设计

注入扰动的类型要尽量丰富，这样才能尽量全面地覆盖系统中可能出现的各种情况。但从成本上面考虑，对所有的扰动进行模拟是不现实的，因此需要对这些扰动进行评估和筛选。具体可采用以下两个筛选思路:

a. 优先考虑和实验假设相关性高的扰动。

b. 优先考虑那些会频繁发生、有代表性或影响显著的扰动。

扰动可以按照不同的作用层级分为以下五个类别:

表 5 混沌工程实验扰动类别

| 扰动类别 | 扰动描述 | 案例 |
| --- | --- | --- |
| 基础硬件资源扰动 | 以各系统运行所需的硬件基础设备为目标的扰动，模拟硬件设备因老化、质量问题和环境因素而发生的故障 | CPU 故障、内存损坏、磁盘写满、硬盘掉盘等 |
| 网络扰动 | 作用于网络连接的扰动，模拟光纤、路由、DNS 的异常造成的网 | 网络抖动、丢包、超时、网卡满、DNS 故障、断网等 |

| | | |
|---|---|---|
| | 络问题 | |
| 系统和中间件扰动 | 作用于系统和中间件资源的扰动，通常是系统或中间件的异常或资源限制 | 操作系统或中间件的崩溃、时钟错误、卡顿等，以及 CPU、内存、磁盘空间等系统资源的占用 |
| 应用扰动 | 作用于实验对象系统内部的扰动 | 连接关闭、进程终止、API 访问故障等 |
| 用户操作扰动 | 用户群体的极端操作行为 | 服务请求激增、异常操作激增、异地访问量激增等 |

来源：中国信息通信研究院，2021 年

### 5. 扰动注入模式的选择

除扰动类型之外，扰动注入的方式也有多种选择的空间。扰动注入的时间点、强度、排列组合方式的不同也会在很大程度上影响混沌工程实验的效果。扰动注入模式一般有以下三种划分维度。

**固定扰动或随机扰动：**扰动注入的随机性通常体现在注入时间点、持续时间、注入扰动的节点、扰动强度和扰动种类这几个方向。随机注入的好处是只要给予充足的时间多次注入，实验可以尽可能全面地覆盖各种突发情况。随机注入的方式适用于对影响范围较小、持续时间较短、恢复成本较低的扰动进行自动化注入。对于影响范围大、持续时间长、实施成本和风险较高的故障类型，则更适合采用设计典型用例，在特定的演练时间按固定的模式注入。固定模式注入的好处是实验的影响范围更为可控，如发生系统失效，更易探究系统失效原因。

**有损扰动或无损扰动：**确定扰动注入强度时，我们应该对扰动可能造成的风险和成本进行评估，在保证实验效果的同时降低损耗。我们可以把不同强度的扰动分为以下几个级别:

a. 无损扰动，且不会造成服务短期或长期的失效。

b. 无损扰动，会造成服务短期的失效（可接受范围内），但会立刻自动恢复。

c. 有损扰动，会造成服务失效，且不可自动恢复。

d. 有损扰动，造成服务失效，造成硬件、操作系统、中间件损坏或客户流失。

有损注入需慎重选择，因为这意味着更多的时间、资金上的投入。很多有损性的扰动更接近真实场景中可能出现的突发事件，具有较高价值，但出于成本的考虑通常会把故障的损耗控制在系统能处理的范围内。

**单一扰动或复合扰动：**单一扰动的注入影响范围更小、更为可控，可以更加精确的定位问题。这种方式更加适用于系统迭代的初期。待单一因素造成的故障被一一排查，可逐步开始进行复合扰动的注入。复合扰动注入可以更有效地发现一些单一扰动不易触发的级联故障，得到对系统的新认知。对于信心较高的系统通常会采用自动化的方式持续注入随机组合的多个无损扰动，这样既可以提升实验效率，又能覆盖各种扰动的排列组合。

### 6. 实验结果预期

在实验设计的过程中需结合对系统的理解对系统受到扰动时可能产生的反应和各项监控指标的变化进行预先评估。这样将有助于发现实验设计中可能存在的问题，提前预估风险，并为应对可能存在的突发情况做好准备。

## （二）混沌工程实验实施

混沌工程实验具备一定的探索性，因而有着比较高的自由度。但在实施过程中，还是有着大致相同的操作流程。明确这些流程对混沌工程试验的规范化是必要的。

### 1. 前期准备工作

混沌工程实验开始前通常会进行一些准备工作，以确保实验在适当的环境中进行。

**确保服务有一定的稳定性基础：**需要确保服务有一定的稳定性，或已经应用了弹性模式，开启了安全壁垒，确保不会因为混沌工程实验造成严重的服务中断。

**准备好完备的监控方式及日志记录：**需要准备好对被测系统进行全面、实时的监控。监控内容不应仅限于系统评估指标，很多看似无关的系统变化对于定位问题都可能会有帮助，这些变化通常很难在实验开始前预估。推荐的做法是尽量使系统保有完备的日志，对事件、流程、关键的数据指标都有实时的记录，这样可提升排查、定位问题的效率。

**准备好处理可能出现的问题：**混沌工程实验开始前需要对系统相关信息和历史的故障案例进行研究，做好应急预案，以应对实验过程中随时可能出现的错误或失效。应确保在系统失效发生时能有相关人员的支持，在短时间内能够排查故障、恢复服务。

**确保在组织内进行了充分沟通：**由于实验过程可能会对系统的运行产生干扰，需在组织内沟通协调，以降低混沌工程实验对其他业务

的影响。

### 2. 实验过程

混沌工程实验相较于传统的测试工作有着更强的不确定性，需要对系统运行状态有更好的把控，并在实验时密切关注系统的反馈，各实验参数也需根据实际情况进行调整，扰动移除后需确保系统恢复正常。

**保持系统正常工作，收集稳态指标：**稳态即目标系统在业务场景下稳定运行时的总体状况，是混沌工程实验的对照组。通常可以通过在无扰动的环境下进行测试得到稳态指标，也可以直接收集系统在生产环境下正常工作时的实时监控信息作为稳态指标，以得到更适用于生产环境的稳态指标，同时避免额外测试。稳态确立的过程需要注意以下两个方面：其一是需要确保系统的各项评估指标在定义稳态的时间范围内保持稳定。其二是收集稳态指标时需要确保这些指标具有实时性，在稳态指标收集的时间粒度上需和混沌工程实验的预期时间跨度相吻合。

**扰动注入：**扰动注入采用的方法通常是运行特定的程序或指令，占用系统资源，或者关闭、干扰特定的组件。在实际实施过程中通常会采用自动化的扰动注入工具。如需进行多组扰动注入实验，建议从影响范围较小的实验组开始着手。在经历一系列小范围的混沌工程实验，对系统有更高的信心后，可以逐步扩大实验范围。

**实验过程监控：**实验运行期间要密切关注相关系统评估指标的变化，关注系统是否出现告警或业务异常。如发生系统失效，可随时终

止实验并采取应对措施。

**实验调整：**在实验过程中，可以根据指标的波动情况，随时调整实验参数，改变扰动的影响范围和强度。如指标波动不明显，可以适当增加实验强度。如系统发生大范围失效，难以定位故障源，则可适度降低实验的强度。

**终止扰动注入：**扰动注入进程达到预定时间后需具备自行终止的能力，如发生较为严重的失效，也可由混沌工程执行人员提前终止扰动注入。扰动注入终止后，需确保没有残留的进程。

**恢复性验证：**系统需要在扰动移除后继续运行一段时间，收集指标并观查系统是否恢复正常工作。

**实验备案：**形成实验报告并归档。需记录实验设计和相关数据，包括实验假设、评估指标变化、扰动注入流程和实验结果。这些实验取得的经验将有助于对系统的改进提供依据，并为后续混沌工程实验的设计提供参考。

### （三）混沌工程实验结果分析

实验的各项操作完成后，需对得到的实验结果进行归纳和分析，这将有助于加深对系统的了解、探究系统的弱点并采取相应的系统稳定性改进措施。

#### 1. 系统稳定性分析

通过比较注入扰动时系统评估指标和稳态时指标的差异，可以对系统的稳定性进行分析。这种分析可以是定性的，如通过功能在注入扰动时是否可用来判断系统是否稳定。但在很多情况下，系统失效的

发生是一个连续的过程：随着扰动强度的增加，系统性能会逐渐下降最终造成系统的不可用，这种情况下量化的稳定性分析往往能更好地反映系统应对压力的情况。通常会计算系统在注入扰动前后的相对性能（P）来反映系统单位时间产生的价值受扰动的影响：

$$P = \frac{E}{E_0}$$

其中 E 为实验组的性能指标，$E_0$ 为稳态时的性能指标。相对性能的数值越高说明扰动对系统的影响越小。另一个重要的稳定性评价指标是系统恢复率，即在扰动移除后系统的恢复情况（R）：

$$R = \frac{E_R}{E}$$

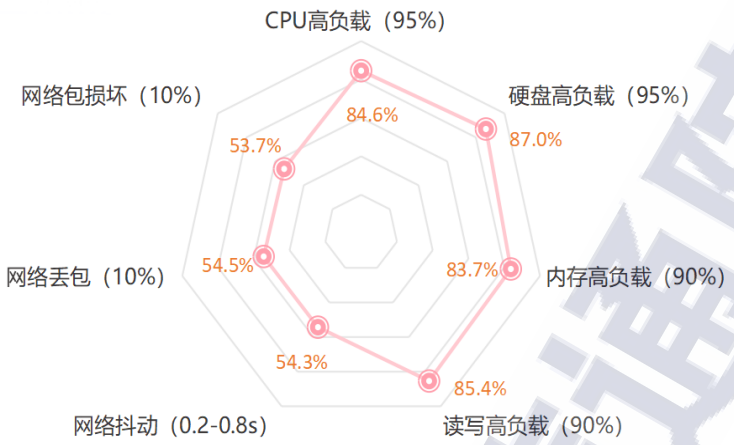
其中 E 为实验组的性能指标，$E_R$ 为移除扰动后进行恢复性验证时的性能指标，系统恢复率越高说明系统恢复越彻底。对于系统资源占用类型的扰动，如 CPU、内存、网络等资源的占用，可以通过计算相对性价比（C）来反映单位系统资源所能达到的性能是否受到扰动的影响：

$$C = \frac{PR_0}{R}$$

其中 R 为实验组的可用系统资源，$R_0$ 为稳态时的可用系统资源，P 为系统在注入扰动前后的相对性能。相对性价比可用于衡量系统对于特定系统资源是否具备冗余性，具备一定冗余性的系统相对性价比是大于 1 的，即在资源移除时性能可以在一定程度上维持。

中国信通院于 2021 年起开展的分布式数据产品稳定性评测项目中即采用上述稳定性分析方法，现已完成多个产品的测评工作，对不

同数据产品进行量化的稳定性评估。




来源：中国信息通信研究院，2021 年

图 5 参与中国信通院产品评测的分布式数据库在不同扰动下的相对性能

## 2. 系统缺陷原因分析

对于稳定性不符合预期的实验组需要考虑以下几个方面：

**对系统中存在的弱点进行分析：**需要根据收集到的系统指标变化趋势和监控日志，分析造成系统失效或性能下降的原因。找到系统弱点，并加以改进。

**对故障应对过程中的不足进行分析：**通过故障复盘，发现故障应对过程中的不足点，完善故障应对流程。

**对系统的扰动承受能力进行评估：**分析扰动对系统性能影响的拐点，以及维持功能可用的零界点，以加深对系统的理解，为后续的混沌工程实验设计和故障应对提供参考。

**对监控告警的有效性进行分析：**需评估实验过程中的告警是否符合预期，评估方向包括：监控告警覆盖度是否足够、监控维度是否正确、告警阈值是否合理、告警是否快速、告警接收人是否正确。根据

评估结果优化无效告警。

**对模块间的依赖关系进行分析：**通过评估注入扰动的模块是否会对其他模块的功能和性能造成影响，可以反映各模块间的依赖关系。根据评估结果对不符合预期的依赖予以改进。

## 三、五个配套措施：战略、人员、文化、风险防范、评估体系

做好混沌工程实践的相关配套措施，是助力混沌工程在机构中融入现有的体系、成功落地实践的关键点。这要求相关机构做好混沌工程实践的战略规划，培养混沌工程实践相关人员，形成混沌工程文化，识别应对潜在的风险并对混沌工程实践做出有效的评估。

### （一）混沌工程实践的战略规划

混沌工程因为需要对现有研发运维体系进行部分调整，对机构有着整体层面的影响，所以实践的战略规划宜采用渐进的方式：先在研发运维部门的个别项目小范围实施，评估其效果，再逐步推广，进行常态化演练的同时使其逐步向标准化、平台化、自动化的方向发展，最终使其成为研发运维体系的一部分。

#### 1. 在测试中使用混沌工程的思想

在混沌工程实践的初期，可以采用投入较少、影响范围较小并具备一定短期收益的实践策略，优先在测试中尝试使用。混沌工程较为保守的应用是通过注入扰动的方式模拟历史故障事件的触发条件，以验证对历史故障的修复情况。这需要测试人员对生产过程中出现过的故障事件进行分析，沉淀为历史故障场景，并引入相应的扰动注入工

具。这种扰动注入测试在工具、人员和实践方式上和更完备的混沌工程实践有一定的连续性，可以看作早期的混沌工程实践。

### 2. 主动设计扰动，发起混沌工程红蓝对抗演练

当有了一定混沌工程实践基础后，可以逐渐采用更主动的方式实践混沌工程。扰动的类型将不止局限于触发历史故障的扰动，部分扰动可根据对系统的理解进行设计。组织内部可以建立混沌工程团队，协同研发运维人员周期性地进行混沌工程红蓝对抗演练：混沌工程团队作为创造问题的一方不断设计混沌工程实验并注入扰动，挑战系统极限；研发运维人员则作为解决问题的一方快速发现故障并修复系统潜在缺陷。一个比较经典的红蓝对抗方式是混沌游戏日。在游戏日中，将由混沌工程团队做决定，向系统中注入扰动，由研发运维方来发现和解决故障并进行故障复盘。游戏日的扰动注入、故障排查、复盘的流程可以进行多轮，可以从较为熟悉的历史故障事件出发，根据新发现的弱点逐步演进，设计新的突发事件。以游戏日方式进行的红蓝对抗不仅可以增强研发运维人员对系统的了解，加深对系统的信心，也使得相关人员更适应承受系统失效的压力。

### 3. 建立混沌工程实验平台

混沌工程实验的平台化有助于提升混沌工程实验的质量和效率。平台可支持模板化的实验设计，并提供稳定的实验场景，使实验规范化，降低人为因素对实验的影响。在生产环境中的实验平台可将用户访问流量的样本发送给实验组，实现最小化爆炸半径 [2]。非生产环境

2 混沌工程实验的爆炸半径即混沌工程实验的影响范围

下的平台通常会包含发起服务请求的模块，从而尽可能真实地模拟业务情况。完善的混沌工程实验平台可允许多个实验同时运行，每个实验可独立进行创建和监控。

#### 4. 混沌工程实验的自动化实施

随着混沌工程实践的推进，依靠人工去完成每个混沌工程实验必然会面临高昂的人力成本和发展的局限性。可以通过工程化能力将实验目标过滤、执行实验、稳态检测、生成实验指标报表、实验结果收集、自动化指标分析等流程串联成自动化的执行流程，减少人力依赖。为了实现实验全流程的自动化，需要对实验案例进行归纳和收集，形成完备的实验集合，平台可自动地从实验集合中按优先级抽取混沌工程实验并实施。实现实验的自动运行也可以更好地支持混沌工程实验的常态化实施。当系统趋于稳定、透明度较高时，混沌工程实验甚至可以在测试或生产环境中不间断的运行，确保开发人员每次提交的改动都能受到混沌工程的检验。混沌工程也将更好地融入系统研发、测试、部署和发布流程中。

### （二）混沌工程实践的人员培养

在有意实践混沌工程的机构中，混沌工程人员的培养可分为两大类：一方面混沌工程对系统的影响是整体层面的，研发运维团队各部门人员均需了解混沌工程理念和基本知识。另一方面，混沌工程是一门实验学科，也需要具有丰富相关知识的专业人员。

#### 1. 整体层面普及混沌工程理念和基本知识

在具备一定混沌工程实践基础的机构中，混沌工程实验通常由专

员负责，但混沌工程实验的实施却可能对整个系统的运作造成影响，需要所有相关人员的沟通协调，以确保不影响其他业务。混沌工程实验所能暴露的问题也是全面的，需要相关研发运维人员配合解决。因此，机构需要确保全体研发运维人员都对混沌工程有一定的了解，消除抵触心理，并对混沌工程实验的实施方式和重要性达成共识。这就要求对全体研发运维人员进行混沌工程相关培训，明确混沌工程的理念和基本知识。

#### 2. 专业层面注重技术和经验的积累

混沌工程是一种实验科学。混沌工程专员在本质上来说和通过实验探索自然现象的科学家是类似的，很多知识并没有现成答案，而是需要在不断的探索发现中积累。因此，混沌工程专员是极其需要实践经验的。除了需要对扰动注入、实验流程编排、系统监控、数据分析的技术有很强的掌握，也需要对系统架构有很深入的理解，这样才能敏锐地发现系统中可能存在的缺陷，提出合理的假设，并有针对性的设计实验。

### （三）混沌工程文化的形成

混沌工程文化是指一个团队在进行混沌工程相关工作时所需的共识和潜意识行为。混沌工程文化是一种建立在相对宽松的企业文化背景之上的，主动面向错误的企业文化。

#### 1. 混沌工程文化的形成需要相对宽松的企业文化背景

混沌工程的实验设计从目标、方式和人员安排上都有很高的自由度，实验过程也具有一定的探索性，往往会发现意料之外的问题，这

就需要一个相对自由、宽松的企业文化。以网飞公司为例，混沌工程之所以能在企业内部得到广泛接纳，和其企业文化密不可分：网飞公司高绩效团队，都具有“认同一致，关系宽松”的特点，这使得每个人在团队里都拥有相同的目标，从而无需花费太多精力来进行过程管理和正式沟通。

### 2. 面向错误、拥抱失败是混沌工程文化的核心内容

混沌工程通过不断试错来提高系统的稳定性，这就要求每个混沌工程相关人员都需要理解“面向错误、拥抱失败”的重要性，将系统故障作为一种学习的手段，以加深对系统的理解。当意识到系统可能存在错误和缺陷时，需要做的是尝试使问题的根源暴露出来，而不是将缺陷隐藏起来，或采用临时性的修补以推迟问题造成的影响。混沌工程文化并不是一蹴而就的，需要经过一段时间的混沌工程实践积累，使相关人员养成通过排查故障积累经验的习惯，在潜移默化中形成混沌工程文化。

## （四）混沌工程实践的潜在风险及应对措施

混沌工程实践的落地需对现有研发运维体系进行部分调整，这一过程中由于技术、人员和管理等不确定因素，存在着诸多潜在风险。在混沌工程实践过程中须注重对这些风险的识别与应对。

### 1. 在生产环境进行混沌工程实验可能造成系统失效

在系统中引入改变通常是引起系统失效的原因之一。因此，在生产环境中引入混沌工程可能存在着造成系统失效的风险，需对其进行技术评估，确保实验的安全无害。如实验风险过高，改进方法是通过

限制爆炸半径，控制潜在的损害，从而避免发生严重事故，也可以首先在准生产或测试环境中运行混沌工程实验，当确保混沌工程实验可以安全进行后，再考虑在生产环境中运行。

### 2. 合规性要求和混沌工程实验发生冲突

合规性要求和混沌工程实验产生冲突是实践混沌工程的另一个潜在风险。很多合规性要求会对混沌工程实验中一些扰动的引入做出限制，阻碍混沌工程实践工作的进行。这些合规性要求通常制定于混沌工程理念被引入之前，并未考虑到通过引入可控的扰动可以防止不可控的大规模事故，这就需要相关人员对合规性要求进行重新评估并做出必要调整。

### 3. 现有系统透明度低或存在无法克服的不稳定性

系统透明度的缺乏将造成混沌工程实验能采用的观测指标不足，这会使得缺陷的识别与定位困难。如系统没有很好的稳定性基础，混沌工程引发的失效会过于频繁，一方面会引起人们对于工作优先级的困惑，对团队士气产生负面影响，另一方面会使故障的定位更为困难。因此，在进行混沌工程实践之前需要对系统的透明度和稳定性基础进行评估。

### 4. 投资回报率评估困难

在管理层面，混沌工程实践的风险因素主要体现在投资回报率的评估上。混沌工程可以防范事故于未然，对于还未造成影响的故障，各方的关注度不高，存在着价值被低估的风险。这将降低各方对混沌工程实践的重视度，影响相关人员的积极性。因此，对相关各方进行

混沌工程实践价值宣贯是极其必要的。

## （五）混沌工程实践的评估体系建立

建立混沌工程实践评估体系可以帮助机构很好地了解混沌工程实践的状况，衡量实践过程中取得的成效，暴露实践中的问题。相关的评估的方向包括混沌工程接纳程度、混沌工程能力、混沌工程价值收益。针对这三个方向，本章节根据业内头部企业的建议，提出了评估方案的基本框架以供参考。实际制定评估方案时需综合考虑机构中的人员组成、业务类型、系统架构等因素。

系统运营者可以根据评估结果有针对性的补齐混沌工程实践过程中的短板：如混沌工程接纳程度不足，可以完善混沌工程人才和文化的培养；如混沌工程能力缺乏，可以有针对性地引入混沌工程工具，建设混沌工程平台，并使其更好地融入现有系统；如混沌工程价值收益过低，则需考虑混沌工程的实验设计是否合理，实验中发现的问题是否能得到有效解决。

### 1. 混沌工程接纳程度评估

可用于评估组织机构混沌工程实践覆盖的广度和深度，评级越高则说明组织机构对混沌工程的接纳程度越高，混沌工程的推进越彻底。

表 6 混沌工程接纳程度评估参考框架

| 混沌工程接纳程度等级 | 1 级 | 2 级 | 3 级 | 4 级 |
|---|---|---|---|---|
| 应用平台种类及个数 | 单个平台 | 单一种类平台 | 少数种类平台 | 多种平台 |
| 应用项目和产品个数 | 单个项目或产品 | 单一种类项目或产品 | 少数种类项目或产品 | 多种项目或产品 |

| | | | | |
|---|---|---|---|---|
| **发现缺陷的影响范围** | 有发现但范围较小 | 有一定的范围，但无法划分 | 有一定的范围，且能划分 | 影响范围较广泛 |
| **发现缺陷的种类** | 单一缺陷 | 单一类别缺陷（如计算、存储、网络） | 多种类别缺陷 | 多种缺陷，缺陷类型全面 |
| **混沌工程实践人员** | 部门内部，兼职 | 部门内部，专员 | 具有混沌工程团队 | 具有混沌工程团队，公司其他人员也积极参与 |
| **混沌工程实践频率** | 偶尔尝试 | 定期进行，周期较长 | 定期进行，周期较短 | 混沌工程实践为日常工作 |
| **混沌工程实践场景** | 单一场景 | 单一类型场景 | 多种类型场景 | 多种类型场景，场景类型全面 |

来源：中国信息通信研究院，2021 年

## 2. 混沌工程能力评估

可用于评估组织机构实践混沌工程实践的能力，主要反映执行混沌工程实践的可行性、有效性和安全性，评级越高则说明组织机构实践混沌工程的能力越强。

表 7 混沌工程能力评估参考框架

| **混沌工程能力等级** | **1 级** | **2 级** | **3 级** | **4 级** | **5 级** |
|---|---|---|---|---|---|
| **架构抵御扰动的能力** | 无抵御扰动的能力 | 一定的冗余性 | 冗余且可扩展 | 已使用可避免级联故障的技术 | 已实现韧性架构 |
| **实验指标设计** | 无系统指标监控 | 实验结果只反映系统状态指标 | 实验结果反映应用的健康状况指标 | 实验结果反映聚合的业务指标 | 可在变量组和对照组之间比较业务指标的差异 |
| **实验环境选择** | 只敢在开发和测试环境中运行实验 | 可在预生产环境中运行实验 | 未在生产环境中，用复制的生产流量来运行实验 | 在生产环境中运行实验 | 包括生产在内的任意环境都可以运行实验 |
| **实验自动化能力** | 全人工流程 | 利用工具进行半自动运行实验 | 自助式创建实验，自动运行实验， | 自动结果分析，自动终止实验 | 全自动的设计、执行和终止实验 |

| | | | | | |
|---|---|---|---|---|---|
| | | | 但需要手动监控和停止实验 | | |
| **实验工具使用** | 无实验工具 | 采用实验工具 | 使用实验框架 | 实验框架和持续发布工具集成 | 并有工具支持交互式的比对实验组和控制组 |
| **扰动注入场景** | 只对实验对象注入一些简单事件，如突发高 CPU 高内存等等 | 可对实验对象进行一些较复杂的扰动注入，如 EC2 实例终止、可用区故障等等 | 对实验对象注入较高级的事件，如网络延迟 | 对变量组引入如服务级别的影响和组合式的异常事件 | 可以注入如对系统的不同使用模式、返回结果和状态的更改等类型的事件 |
| **终止扰动注入能力** | 扰动无法独立终止 | 人为干预，长时间可终止 | 人为干预可终止 | 可定时终止 | 可依据触发条件自动终止 |
| **故障监控能力** | 无法监控 | 能获取到少量数据信息 | 可人为搭建监控 | 自带监控仪表盘 | 自带监控仪表盘和告警能力 |
| **定位问题能力** | 无法定位 | 可人工定位 | 可自动定位 | 可自动精准定位 | 自动精准定位，提供改进方式 |
| **环境恢复能力** | 无法恢复正常环境 | 可手动恢复环境 | 可半自动恢复环境 | 部分可自动恢复环境 | 韧性架构自动恢复 |
| **实验结果整理** | 没有生成的实验结果，需要人工整理判断 | 可通过实验工具的到实验结果，需要人工整理、分析和解读 | 可通过实验工具持续收集实验结果，但需要人工分析和解读 | 可通过实验工具持续收集实验结果和报告，并完成简单的故障原因分析 | 实验结果可预测收入损失、容量规划、区分出不同服务实际的关键程度 |

来源：中国信息通信研究院，2021 年

## 3. 混沌工程价值收益评估

可用于评估组织机构实践混沌工程后所产生的收益，主要考察通过混沌工程实践是否能够发现并解决系统中的问题，是否能够对监控、告警进行优化，并提升机构故障应对的能力。

表 8 混沌工程价值收益评估参考框架

| 混沌工程价值收益等级 | 1 级 | 2 级 | 3 级 | 4 级 |
|---|---|---|---|---|
| 解决问题的应急效率（问题处理时间/解决问题需要的人员数） | 低 | 中 | 较高 | 高 |
| 缺陷复发率 | 高 | 较高 | 低 | 趋近于 0 |
| 生产过程中单位时间内缺陷发现数 | 多 | 较多 | 少 | 趋近于 0 |
| 修复缺陷的严重程度 | 较轻 | 轻 | 中等 | 严重 |
| 监控告警时间（发现问题所需时间） | 长 | 较长 | 较短 | 短 |
| 系统透明度 | 低 | 较低 | 较高 | 高 |
| 混沌工程实验效率 | 低 | 较低 | 较高 | 高 |

来源：中国信息通信研究院，2021 年

## 四、两个延伸保障：加强架构和制度保障

稳定性是系统抵抗扰动并维持稳态的能力，当通过混沌工程发现了系统稳定性缺陷时，需要根据实际情况给出对应的解决方案。高层次的技术缺陷是一些架构上的隐患，需针对性地采用一些韧性设计对稳定性缺陷进行改进。低层次的技术缺陷体现在对于一些边界条件和极端情况缺乏考虑或者编码失误，这些缺陷占比虽高但较易解决。如同类型的低层次技术缺陷反复出现，则需评估是否需要在制度纪律上对研发运维过程进行更好的管控。

### （一）提升系统架构的韧性

对于系统架构中发现的稳定性痛点，可以评估以下韧性架构是否适用。

表 9 基于混沌工程实验的系统架构优化方向

| 架构优化方向 | 描述 | 适用情况 | 架构使用案例 |
|---|---|---|---|
| 冗余设计 | 对资源留出安全的余量 | 系统的正常工作极易受到系统资源限制等扰动的影响 | 重要的数据库项目建设中可以采用异地多活，确保服务不会轻易中断 |
| 无状态设计 | 服务单元只涉及逻辑处 | 混沌实验中故障的原因经常被定位在 | Web 服务器将状态保留在客户端，从而使客户端的多次请求不必访 |

| | | | |
|---|---|---|---|
| | 理而不存储状态，方便服务崩溃时业务的迁移 | 某个模块由承压超过阈值而崩溃 | 问同一台服务器，确保服务的稳定 |
| 故障隔离 | 将故障的影响限制在较小的范围内，避免级联故障的发生 | 扰动注入的影响范围大于预期 | 消息中间件在推送消息时，会启动调节策略，将没有响应的消费节点剔除，避免损失更多的系统资源 |
| 过载保护 | 在服务请求超过服务能力时，适当减少服务接收的比率 | 用户请求激增、容量超额等实验场景易引发全面的服务受损 | 在系统资源不足时采取限制流量（限流）或终止服务（熔断）等措施 |
| 有损服务 | 在服务能力不够的异常情况，系统可以有所取舍 | 用户请求激增、容量超额等实验场景易引发较严重的服务受损 | 直播业务在带宽有限的情况下，会降低码率减少清晰度，而不应该拒绝服务 |
| 去关键路径、关键节点 | 关键路径或节点是系统稳定性短板，应尽量避免 | 在某个链路或节点进行扰动注入对系统整体造成了较为严重的影响 | 军用系统中常常采用去中心化的设计，避免关键节点损失对整个系统造成重大影响 |
| 负载均衡 | 尽量平均地分配系统所受到的压力，分散压力对系统的影响 | 在混沌工程实验中限制单一服务实例的服务能力，其工作并未被其他服务实例分担 | Kubernetes 提供多种负载均衡方式，使系统资源可以按不同的需求充分的利用 |

来源：中国信息通信研究院，2021 年

## （二）加强研发运维过程中的制度保障

如混沌工程实践的过程中发现大量低层次缺陷，通常需要评估研发运维过程中制度纪律是否有待改进。通过制度纪律去规范操作和行为被证明是保障稳定性并减少出错发生的有效方式。但需要注意的是纪律只是划出产品质量的底线，只能解决低层次的稳定性问题。

### 1. 研发过程管控

研发过程中可采取一系列措施提高代码质量，减少低层次失误的产生。

**制定详细的编码规范：**编码规范可以有效改善代码的可读性、严谨性，保持开发风格的统一，便于在团队协作中理解他人的工作内容，减少由于理解不同造成的失误。

**设置代码提交门禁：**代码门禁是一项代码质量保障措施，目的是要求开发人员提交的代码必须满足一些要求才能合入代码仓库。门禁要求可包括：代码需编译成功、通过静态代码扫描，通过动态代码分析等。

**进行代码评审：**由人工评审的方式来减少开发过程中的失误，避免不合理的设计，提早发现缺陷。

### 2. 测试过程管控

确保测试环节的完整性，单元测试、压力测试、回归测试等环节不能缺失。同时需注意测试用例的覆盖率需尽量全面，避免遗漏。

### 3. 发布过程管控

尽量采用灰度发布的方式，确保有完善的回滚机制和应急预案。如发生故障，需对故障进行复盘分析。

## 五、混沌工程发展趋势

一是随着混沌工程概念进一步普及，将有更多的机构实践混沌工程。二是技术的进步将为混沌工程实践赋能，使实践过程更加自动化、智能化。三是混沌工程将随着物联网、区块链等新兴技术的普及而拥有更广的应用范围。

## （一）产业环境和政策导向加速混沌工程实践落地

随着社会各行业现代化演进的逐渐成熟，各组织机构在运行维护上的支出将逐渐超过开发建设成为大头。提高稳定性就意味着减少突发情况对主体带来的影响，增强主体对外界环境变化的抵抗能力，在降低运维成本的同时也使得运维成本更可控。组织机构层面的稳定性在很大程度上取决于其信息基础设施的稳定性，9 月 1 日正式实施的我国关键信息基础设施安全保护重要法规《关键信息基础设施安全保护条例》对关键信息基础设施的相关保障工作提出了更高的要求。混沌工程作为主动发现系统缺陷的手段，在保障 IT 系统的稳定性方面有着无可替代的作用，必将有越来越多的机构实践混沌工程。Gartner 预计，到 2023 年 40% 的机构将把混沌工程实践作为研发运维的一部分，这将使得故障停机时间减少 20%。

## （二）智能技术推动混沌工程实践更加自动化

混沌工程相关技术将进一步发展。经过深入研究，我们总结出了几个发展方向，一些头部企业目前已经开始了相关的研究:

### 1. 智能化弱点识别

智能化弱点识别技术是混沌工程和人工智能、机器学习等领域相结合的产物，目前已有企业进行了相关的探索工作。通过历史记录中系统在注入故障时指标变化和定位到的弱点来训练机器学习模型，使得模型可以通过指标变化来自动识别弱点。这将有助于解决混沌工程在结果分析和故障恢复环节的自动化问题。

### 2. 可视化交互式演练平台

随着混沌工程的发展，混沌工程实验平台的功能将进一步完善。各平台研发团队将在基本的实验管理、扰动注入、指标观测等功能的基础上，逐步完善演练场景编排、指标分析、演练场景模型管理等功能。演练平台的可视化、自动化和可交互性也将进一步提高。

### 3. 自动化依赖强度分析

传统的静态代码分析可以形成系统各模块的依赖关系网，但无法定量的分析依各模块的赖强度。量化的依赖强度分析可以通过向某一模块注入扰动并观测其他模块受影响的程度来实现。通过自动化的依赖强度分析，可以构建更完善的量化依赖网络，助力系统架构的进一步完善。

### 4. 用于故障模拟的自动化硬件设施

可自动进行故障插入的硬件设施最早出现在汽车制造行业中，用于测试电路短路对车辆行驶造成的影响。目前混沌工程中扰动注入的自动化主要是在系统层以上的层级。在模拟内存损坏、硬盘掉盘、磁道损坏、接触不良等底层硬件故障时，通常需要进行人工的操作，如拔掉硬盘、网线等操作。可进行自动化控制的混沌工程硬件将有助于解决这一问题。

## （三）数字技术的推广应用将带动混沌工程推广落地

目前，各行业正利用物联网、区块链、分布式数据库等数字技术，推动信息化与实体经济深度融合，促进业内数字化转型不断深化、细

化。而数字技术的应用，也迫使各行业 IT 系统直接面临系统稳定性风险的跃迁。所以混沌工程将伴随着这类数字技术，逐渐推广落地。

### 1. 物联网

物联网是将各种信息传感设备通过互联网连接形成的分布式网络。相对于机房部署的业务系统而言，物联网所处的网络环境更为复杂。设备会频繁地出现掉线、回复短路等情况，如不能很好的处理这些问题将严重影响用户体验。在物联网设备研发过程中，通过混沌工程来模拟物联网的连接问题，可以很好地对设备的抗扰动能力做出评估。

### 2. 区块链

区块链是由参与者通过网络连接和内部算法来创建并维护的分布式去中心化账本。区块链技术的发展过程中暴露出的一些脆弱性问题逐渐引起关注，例如以太坊(Ethereum)的 DAO 事件。混沌工程可以更早的暴露这些风险并在早期寻求解决方案。目前腾讯等企业已经开始进行区块链场景下的混沌工程实践，开源的混沌工程实验工具包 Chaos Toolkit 也开放了对区块链场景的支持。

### 3. 分布式数据库

作为 IT 系统的重要基础组件，数据库的稳定性至关重要。对于分布式数据库来说，具备节点和网络故障的抵抗能力和自愈能力是基本的要求。MongoDB 的研发团队就将混沌工程作为每一次发布过程中必须执行的环节，并建立了供内部使用的混沌工程平台 Evergreen，可进行的混沌工程实验包括网络分区、系统时钟漂移和节点崩溃等。

# 附录

## （一）业内混沌工程工具一览

混沌工程工具的种类繁多，其适用的环境各有不同，包括云平台环境、传统操作系统层面和特定应用程序等等。注入扰动的种类也各有侧重。86%的混沌工程工具都是以开源项目的形式研发并发布的。

表 10 混沌工程工具总结

| 工具名称 | 最新版本 | 项目维护状态 | 主要构建语言 | 涉及场景 | 特定依赖 |
|---|---|---|---|---|---|
| Chaos Monkey | 2.0.2 | 停滞 | Go | 终止 EC2 实例 | Spinnaker |
| Simian Army | 2.5.3 | 废弃 | Java | 终止 EC2 实例，阻断网络流量，卸载磁盘卷，CPU/IO/磁盘空间突发过高，终止进程，路由失败，网络丢包，DynamoDB 故障 | 无 |
| orchestrator | 3.1.1 | 活跃 | Go | 纯 MySQL 集群故障场景 | 无 |
| kube-monkey | 0.3.0 | 停滞 | Go | 终止 K8s Pods | 依赖于 K8s 集群 |
| chaostoolkit | 1.2.0 | 活跃 | Python | 实验框架，可集成多个 IaaS 或 PaaS 平台，可使用多个扰动注入工具定制场景，可与多个监控平台合作观测和记录指标信息 | 通过插件形式支持多个 IaaS、PaaS，包括 AWS/Azure/Google/K8s |
| PowerfulSeal | 2.2.0 | 活跃 | Python | 终止 K8s、Pods，终止容器，终止虚拟机 | 支持 OpenStack/AWS/本地机器 |
| toxiproxy | 2.1.4 | 活跃 | Go | 网络代理故障，网络故障 | 无 |
| Pumba | 0.6.4 | 活跃 | Go | 停删容器，暂停容器内进程，网络延迟，网络丢包，网络带宽限流 | 依赖 Docker |

| | | | | | |
|---|---|---|---|---|---|
| blockade | 0.4.0 | 停滞 | Python | 终止容器，网络终端，网络延迟，网络丢包，网络分区 | 依赖 Docker |
| chaos-lambda | 0.3.0 | 停滞 | Python | 终止 EC2 示例 | 依赖于 Lambda |
| namazu | 0.2.1 | 停滞 | Go | 文件系统故障，网络故障，Java 功能调用故障 | 针对类 Zookeeper 分布式系统 |
| Chaos Mesh | 2.0 | 活跃 | Go | 实验框架，支持系统资源、网络、应用层面等多种故障的注入 | 依赖于 K8s 集群 |
| byte-monkey | 1.0.0 | 停滞 | Java | 异常处理，应用延迟 | 无 |
| Chaos Blade | 0.4.0 | 活跃 | Go | 实验框架，支持系统资源、网络、应用层面等多种故障的注入 | 无 |

来源：中国信息通信研究院，2021 年

## （二）混沌工程平台简要介绍

混沌工程平台在结构上有较高的一致性，主要由用户界面、任务调度模块、扰动注入介质、监控告警系统、测试模型库五部分组成。

### 1. 用户界面

用户界面提供各类混沌工程实验任务的编排和配置服务，借助演练流程编排面板，用户可以便捷地管理各类混沌工程实验任务；混沌工程实验开始实施后，用户可通过任务进度条、服务器指标展示图等实时查看实验进度和系统指标情况；混沌工程实验执行结束后，用户界面会展示相关指标，生成混沌工程实验报告。

### 2. 任务调度模块

任务调度模块负责用户界面和扰动注入介质之间的交互，核心功

能是实现混沌工程实验任务的批量下发和调度，该模块可以批量下发各种类型的混沌工程实验。

### 3. 扰动注入介质

扰动注入介质负责接收任务调度模块下发的扰动注入任务，实现相应的扰动注入事件，并反馈扰动注入任务的执行状态。

### 4. 监控告警系统

监控告警系统负责记录和管理系统产生的所有数据，生成告警和相关统计并反馈给用户界面。

### 5. 测试模型库

测试模型库包含混沌工程专员根据平时混沌实验总结得到的测试模型，基于测试模型库用户可以根据演练场景自动关联对应的扰动注入事件，并为用户提供一键生成混沌实验流程的能力。


来源：中国信息通信研究院，2021 年

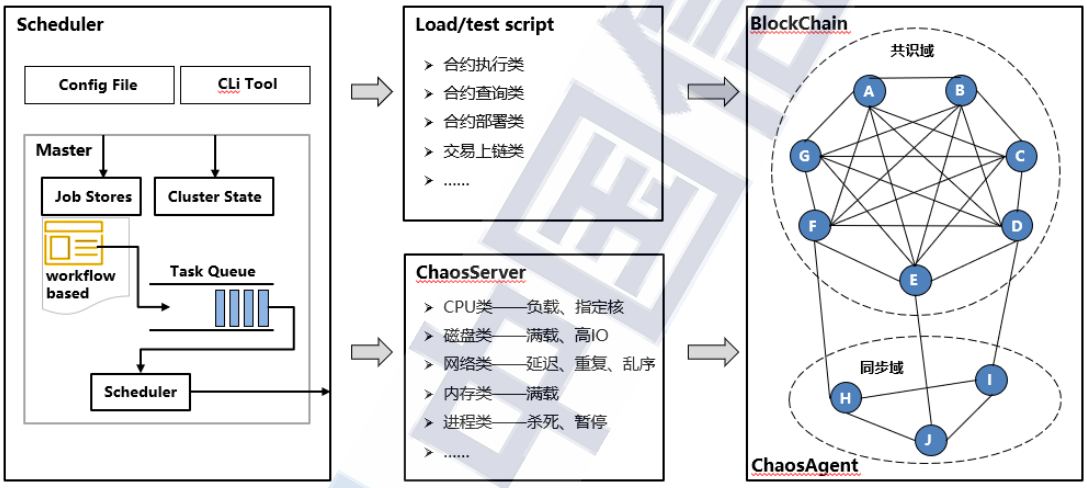
图 6 混沌工程平台架构

## （三）混沌工程实践案例

本节选取两个有代表性的混沌工程实践案例，以供参考。

## 1. 腾讯区块链项目混沌工程实践

区块链服务建立在分布式系统之上，可能遇到的故障非常多，如果区块链算法不能很好的处理这些异常，那么服务的稳定性将遭到挑战，其后果也不堪设想。出于此考虑，腾讯区块链团队建立了混沌工程实施框架，对区块链产品进行自动化的网络分区实验，容错场景验证和告警有效性验证。




来源：腾讯区块链团队，2021 年

图 7 腾讯区块链团队混沌工程实施框架

整个实施框架分为以下几部分：

**被测区块链网络区域：**搭建好的区块链网络分为共识域和同步域，每个区块链节点上都会部署相应的故障注入模块。

**压测脚本：**用来对被测的区块链网络施加稳定的基准流量，这些流量用例包括合约执行、交易上链、交易查询等。

**混沌执行服务器：**负责接收指令请求，向具体的节点施加具体的异常场景，包括 CPU、磁盘、网络、进程异常等。

**混沌场景调度器：**在特定的时间点向某个节点施加异常状态，以工作流的形式放入任务队列中，然后调度器按时发放任务。

混沌工程在腾讯区块链团队的实施以渐进的方式进行，推进方式由固定时间、固定故障类型向不定期、随机故障类型转变。与此同时，混沌工程团队也致力于系统透明度的完善并积极排查无效告警。依托自动化的混沌工程实施框架，混沌工程团队逐步形成了成熟的故障场景模拟和应对能力，可以实现分钟级别的问题发现和告警能力，累计发现系统隐患 50 余个。腾讯混沌工程实践案例的成功得益于混沌工程实验的高度自动化实施以及系统透明度的提升：同手动进行混沌工程实验相比，实验的自动化实施使得实验实施时间从平均 25 分钟缩短到现在的 5 分钟；系统透明度的提升使得故障定位的时间从原有的平均 10 分钟缩短到现在的平均 3 分钟，这些改进极大地提升了混沌工程实验的效率。

## 2. 华为云稳定性门禁验收过程中的混沌工程实践

华为云的混沌工程的成功实践得益于公司 30 多年的电信级产品研发经验以及丰富的云服务的典型故障场景应对经验。华为云系统保障团队通过长时间的持续积累，制定出了华为云故障模式库，为混沌工程的实践打下了坚实的基础。

| 端到端故障模式库（系统可靠性） | | |
|---|---|---|
| I层 | II层 | 故障模式 |
| 设备和电源(72) | 交换机(24) | 主用交换机故障(3)　备用交换机故障　交换机端口故障(8)　交换机链路故障(3)　交换机工作异常(10) |
| | 路由器(1) | 主用路由器故障(1)　备用路由器故障 |
| | 防火墙(9) | 主用防火墙故障　备用防火墙故障(1)　防火墙工作异常(8) |
| | 物理链路(20) | 物理链路闪断(3)　误码率高(4)　光模块故障(8)　专线故障(5) |
| | 机房电源(18) | 单路故障或供电切换(18)　供电不足(*) |
| 机房环境(88) | 机房环境(34) | 制冷故障(10)　机房掉电(13)　机柜掉电(11) |
| | 网络平面(54) | 网络区域故障(4)　网络平面故障(7)　网络亚健康(16)　网络风暴(3)　网络地址冲突(14)　IP不可用(9)　路由负载不均衡(1) |

来源：华为云系统保障团队，2021 年

图 8 华为云故障模式库内容概览

基于华为云端到端的故障模式库，系统保障团队制定出了稳定性评估基线。该基线是云服务稳定性的最低要求，为云服务的稳定性门禁验收工作提供依据。稳定性测试工程师在测试验收环节会严格按照基线要求进行验收：通过混沌工程的方式对故障模式库中存有的主要的故障进行编排并注入到待发布的服务中，即可观测并验证云服务的稳定性是否达到稳定性评估基线，如果服务无法满足稳定性评估基线则不能通过验收，也无法上线。这种基于混沌工程的稳定性验收工作很好的保证了华为云服务的稳定性。门禁验收过程可通过华为云自研的智能测评平台进行自动化实施。该平台在公司内部被广泛使用，支撑开发测试人员和站点可靠性工程师进行测试、演练和可靠性门禁验收。

中国信息通信研究院 云计算与大数据研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：13011807607

传真：010-62304980

网址：www.caict.ac.cn