分布式系统稳定性实验室
Distributed System Stability Lab

云计算与大数据研究所

信息系统稳定性保障能力建设指南（1.0）

（2022 年）

分布式系统稳定性实验室
中国信息通信研究院云计算与大数据研究所

2022 年 3 月

版权声明

编写委员会

主要编写单位:

中国信息通信研究院云计算与大数据研究所、杭州数列网络科技有限责任公司、蚂蚁科技集团股份有限公司、腾讯云计算（北京）有限责任公司、博睿宏远数据科技股份有限公司、上海新炬网络信息技术股份有限公司、中国光大银行、中信银行、国家电网有限公司客户服务中心、浙江大学信息技术中心教学支持部

参与编写单位:

中国移动通讯集团浙江有限公司、中国农业银行、深圳市杉岩数据技术有限公司、上海驻云信息科技有限公司、中国工商银行数据智能中心、数字广东网络建设有限公司、平安科技（深圳）有限公司、北京城市网邻信息技术有限公司、广州虎牙信息科技有限公司、中国建设银行、上海哔哩哔哩科技有限公司、顺丰科技技术集团、北京趣拿软件科技有限公司、中国联合网络通信有限公司软件研究院、中移（苏州）软件技术有限公司、中移动金融科技有限公司、中国人寿保险股份有限公司、上海德邦物流有限公司

编写组主要成员:

王超伦、魏凯、姜春宇、马鹏玮、王卓、杨佳星、杨德华、陆学慧、陈志宏、钟圣荣、黎进财、古韬、张亮、刘敬仁、林志、田凌翔、刘颢、张俊峰、王紫博、蒋晓君、刘光宇、刘帅、程永新、宋

辉、梁铭图、翁建敏、张勇、黄国标、郭岳、叶晓龙、史军艇、赵隆兵、张观石、张才俊、李琦、杜俊、唐守忠、周越德、花瑞、王丽华、李祖金、张志海、李俊谦、李钦、罗新良、邹鹤良、陈奇、李波、戈先武、刘元、陈亮、张宇燕、李卓、陈阳、王志广、朱仕智、李瑞、吴天昊、李明亮、孙小霞、朱刘江、陈明、毕海波、张泉、陈尚荣

前　言

随着各领域数字化转型的推进，信息系统的应用范围不断扩大、承载业务愈发关键，用户的高频访问成为常态。面对使用需求的不断增长，大多数信息系统通过分布式架构改造、DevOps 体系建设、大量引入开源技术来不断突破自身处理能力上限，这些措施引入导致了信息系统架构复杂性呈指数上升，显著增加了稳定性风险。2021 年，谷歌、亚马逊、微软、滴滴、特斯拉等大型互联网公司均发生了大范围宕机事件，在对自身造成损失的同时，也严重影响了全球范围内用户的正常生产生活。信息系统的稳定性也受到国家高度重视，2021 年 9 月 1 日正式实施的《关键信息基础设施安全保护条例》中对我国关键信息基础设施的稳定性保障工作提出了明确要求。

信息系统稳定性保障能力建设工作已有多年发展历史，然而随着分布式架构的普及、大流量高频访问的常态化、关基等政策的颁布，系统稳定性保障工作进入了新的阶段，需要在理论和实践两方面均进行大量优化。在此背景下，中国信息通信研究院信息系统稳定性实验室组织业内头部企业编写本指南，梳理并总结了新阶段下信息系统稳定性保障能力建设工作的相关背景、基本原则、关键要素、核心能力以及评价体系，探讨了稳定性保障工作的未来发展趋势。本指南旨在总结最佳实践，形成方法论，帮助各行业、各机构完善信息系统稳定性保障体系，助力数字化转型“又快又稳”。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、信息系统稳定保障体系概述

信息系统的稳定性即信息系统平稳连续运行的能力。一个稳定的系统需要具备功能的连续性并维持一定的性能，长久而不间断地提供服务。业界通常采用平均无故障时间（MTTF）和平均维修时间（MTTR）来定义稳定性[1]，稳定性高的系统需具备较低的故障频率和较短的故障维修时间。由于信息系统具备本质上的复杂性，系统稳定性保障能力的建设也是一个复杂的过程，涉及到系统生命周期的各个方面，任何环节出现短板即可能造成隐患。因此稳定性保障工作需要建立一个完整的配套体系。本章梳理了信息系统稳定性保障工作的发展历程，总结了新阶段下信息系统稳定性面临的挑战，并提出了当前阶段下信息系统稳定性保障体系的基本框架。

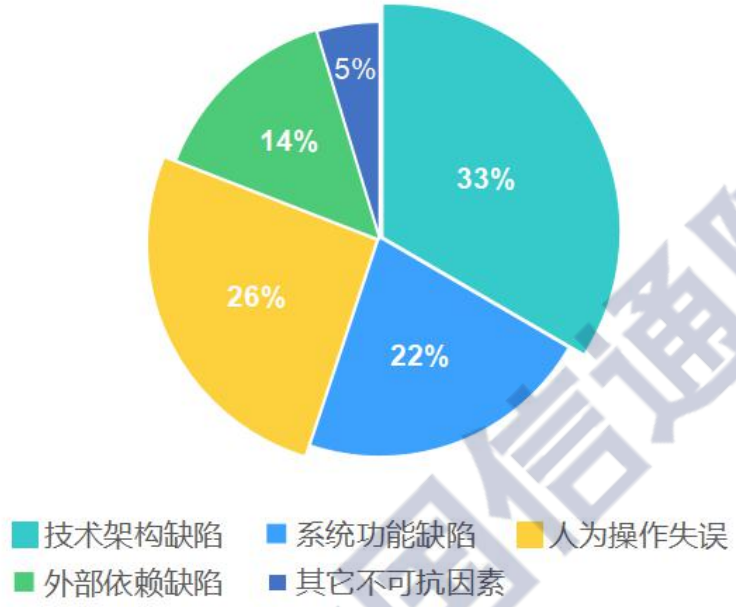
（一）复杂性是信息系统稳定性隐患的原罪

信息系统是以处理信息为目的的计算机系统，包括计算机硬件和软件，被用于处理逻辑或业务规则比较复杂的问题。信息系统本质上是复杂的系统，因此也具备复杂系统所具备的两个特点：一是脆弱性，即较易受到变更和环境差异的影响；二是模糊性，即认知成本高，难以通过单一组件的属性预计整体事态的发展。

这种复杂性作用于信息系统开发、测试、运维的各个方面，导致大量稳定性缺陷的引入。这些缺陷的根因可分为系统技术架构缺陷、系统功能缺陷、人为操作失误、外部依赖错误、其它不可抗因素（如机房电路老化）等部分。通过调研多家政务、金融、互联网、电信领

[1] 系统稳定性（S）可由平均无故障时间（MTTF）与平均维修时间（MTTR）计算得到：S=MTTF/（MTTF+MTTR），云服务等产品通常在 SLA（Service Level Agreement）中定义

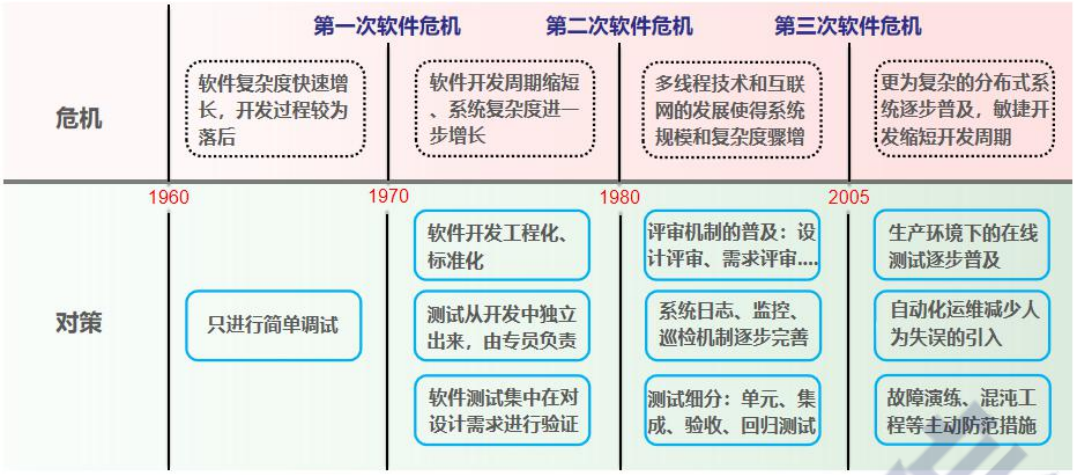
域的头部企业的核心业务系统，我们汇总出各类故障根因所占的比例，如图 1 所示，其中技术架构缺陷占比和人为操作失误占比较高，这两者也是与信息系统的脆弱性和模糊性直接相关的。



来源：中国信息通信研究院，2022 年

图 1 信息系统故障根因所占比例统计

自 1960 年以来，信息系统中的软件部分从最初的单节点、单线程向分布式、微服务的方向不断演进，相较于硬件而言，软件系统复杂度的增长更为明显，软件系统逐步成为了稳定性事故的重灾区。历史上多次爆发的软件危机也给信息化从业人员敲响了警钟，推动着各方建设更加完善的稳定性保障体系。



来源：中国信息通信研究院，2021 年

图 2 系统稳定性危机与对策发展时间线

（二）IT 行业高速发展对系统稳定性带来新的挑战

随着互联网技术的普及，IT 产业在 21 世纪迎来了快速发展的时期，信息系统在规模、研发运维模式、技术架构、用户群体上均发生重大变化：

系统分布式化后故障概率增大：由于单机的性能瓶颈，越来越多的信息系统采用分布式架构，一方面，随着硬件数量的增加，底层硬件和网络设备故障的概率随之增加。另一方面，分布式架构通常包含多种异构的硬件设备，这些设备的损耗、替换速率存在差异，这种异构性也增加了硬件设备发生故障的概率。

开发体系的演化引入新的风险：由于系统的规模逐步增大，开发体系也在不断迭代，新开发体系的应用将引入新的风险。一方面，大型系统需要多个团队之间进行协作，这增加了缺陷引入的可能；另一方面，敏捷开发逐渐成为主流，在缩短开发周期的同时也导致一些需要较长时间才能暴露的缺陷引入到系统中。

大量新兴技术的采用带来稳定性隐患：近十余年来大数据、云计算、人工智能、元宇宙领域新兴技术迅速发展。各行业的 IT 系统迭代迅速，以更快地吸纳这些新兴技术，率先占领相关市场。系统长期运行的稳定性往往会被忽视，欠下技术债。

疫情促使大量的业务转为线上：随着国际上疫情的持续蔓延，国内疫情防控的常态化，各线上系统的日活用户都有显著的增长，在线办公、消费、娱乐逐渐成为主流，这将对各类信息系统稳定性带来很大挑战，尤其是还不具备长期稳定性保障能力的传统行业。

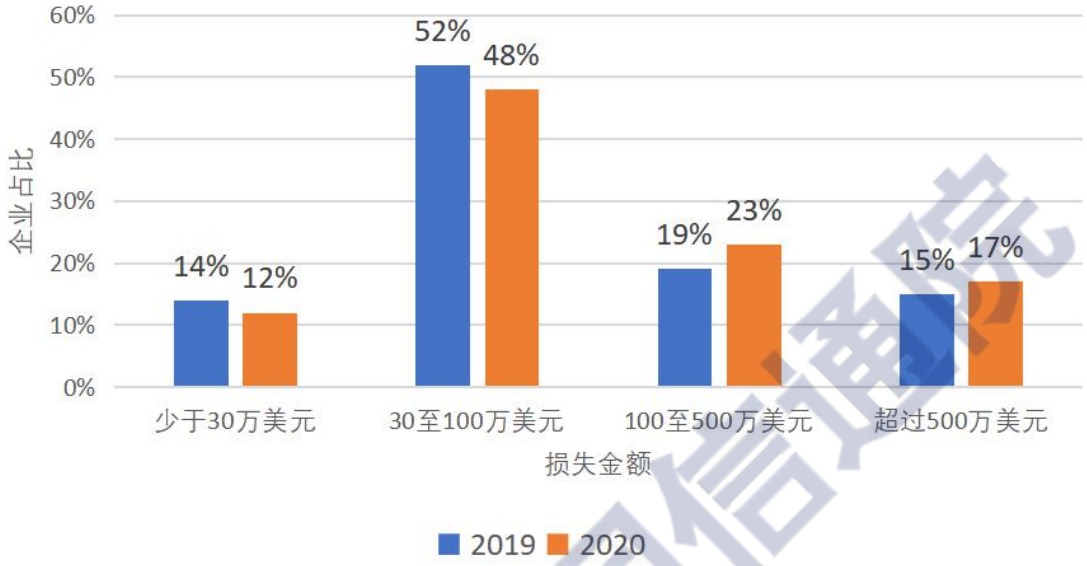
随着信息系统成为各行业基础设施，信息系统故障的影响范围越来越广，其后果也越来越严重。表 1 统计了 2021 年全球重大的信息系统事故，其中不乏亚马逊、特斯拉、Facebook 等行业巨头。

表 1 2021 年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
亚玛逊	2021 年 12 月	约 3 小时	全球亚玛逊云计算服务	数据中心及网络连接问题
特斯拉	2021 年 11 月	约 5 小时	特斯拉 App 全球范围服务中断	配置错误导致网络流量过载
Facebook	2021 年 10 月	约 7 小时	Facebook 及旗下 Messenger、Instagram、WhatsApp 等多个服务	运维操作失误
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

来源：中国信息通信研究院，2021 年

行业分析机构 Statista 对关键的线上服务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%。



数据来源：Statista，2020 年

图 3 企业服务中断每小时造成的损失统计

（三）数字化时代亟需建立新的信息系统稳定性保障体系

近年来，各行业数字化转型加速，信息系统稳定性面临诸多新挑战，时代呼吁新的系统稳定性保障体系。政策方面，2021 年 9 月 1 日正式实施的《关键信息基础设施安全保护条例》对我国关键信息基础设施的稳定性保障工作提出了更高的要求。技术方面，AIOps、混沌工程、全链路压测等新技术的出现也为稳定性保障工作提供了新的思路。然而由于稳定性保障相关工作内容较为零散，涉及信息系统生命周期的各个方面，稳定性保障体系的全面建设仍仅限于头部企业。为促进我国信息技术产业平稳健康发展，信通院分布式系统稳定性实

验室汇总并梳理了互联网、电信、金融等领域内头部机构的稳定性保障实践经验，首次提出了信息系统稳定性保障体系。如图 4 所示，该体系共包含“两个总体原则、三个关键要素、四个核心能力、五项重要工作”，本小节首先对稳定性保障体系中的各项进行总括性介绍，后续各章节将针对各部分内容进行详细叙述。

来源：中国信息通信研究院，2022 年

图 4 信息系统稳定性保障体系基本框架

总体原则：平衡取舍、积极防御。系统稳定性保障工作须统筹考虑用户体验、效率、成本等因素，稳定性保障工作建议采取积极防御的原则，以应急演练等方式促进保障体系的完善。

关键要素：人员、管理、技术是稳定性保障的三个关键要素。稳定性是一个整体层面的属性，人员、管理、技术三个方面中的任何一项存在短板都会造成稳定性隐患的引入。

核心能力：故障预防、识别、响应、优化是四项核心能力。稳定性保障能力的建设具备长效性，是对故障预防、响应、处置和改进能力的不断提升，需融入机构的日常活动中。

重要工作：关键时期稳定性保障需要主抓团队组织，进行全面梳理，做好预案准备，做到事中协同、事后复盘。关键时间节点意味着更加频繁的用户访问，系统较易因为压力出现问题，潜在的逻辑漏洞更易暴露，宕机损失也更加巨大。因此通常会在关键时期开展重点保障工作。

二、两个总体原则：平衡取舍、积极防御

系统稳定性保障工作须满足平衡取舍、积极防御的总体原则。一方面，保障系统平稳运行是和系统生命周期的各项工作息息相关的；另一方面，也应该兼顾被动的故障应对和主动演练。

（一）原则一：平衡用户体验、效率、成本的关系

系统在建设和维护的过程中即需对稳定性做出相关要求和考虑，如韧性架构的使用、规范化的开发测试流程、建立完备的监控告警等。这些稳定性方面的考虑并不是没有代价的。需要综合考虑用户体验、效率、成本等因素，将稳定性保障工作的投入产出比最大化。

用户体验：稳定性保障工作需关注用户的使用体验。在理想情况下，系统的稳定程度是和用户体验的好坏正相关的。但由于系统资源的限制，很多时候为了确保系统能在高峰流量下保持稳定，需对系统的一些非核心功能做出权衡取舍，提供有损服务。例如，某电商平台于大型促销活动期间会采用暂时关闭退货、售后等非核心功能，确保核心业务的连续性具备足够的系统资源保障。

效率：从流程管控上来看，业务效率的提升一定程度上会影响到稳定性，反之对稳定性要求过高也会带来对业务效率的降低。效率和

稳定性兼顾是一个困难的问题，需进行一定的取舍。通常成熟的业务会更偏向于稳定性，而新业务更偏向于效率。

成本：一方面，增加稳定性保障工作的力度意味着需要投入更多的资源。另一方面，系统稳定性的提升也会减少系统故障造成的客户流失及品牌影响。对稳定性保障工作的投入并非越大越好，即使将人为引起的故障控制在最低，系统依然有一定的概率由于不可控外因的影响造成宕机，因此过度的投入能带来的收益是有限的。对稳定性保障工作的投入需结合投入产出比进行权衡。

（二）原则二：通过演练促进保障体系的建设

稳定性保障工作需采取积极防御，以演促防的原则，对突发故障进行应对，也需通过应急演练、全链路压测等方式完善防御机制。

故障应对：对生产环境中突发异常状况的应急处理在稳定性保障工作中具备着最高的优先级。系统故障时间越长，造成的损失越大，因此故障应对的首要工作就是尽快恢复系统。然而仅依靠故障发生时的应急处理难以维持系统的长期稳定，一是由于故障应对需在有限的时间内完成，对系统的应急修复往往是临时且有局限性的，缺乏长期、全局的考虑；二是由于突发故障的类型和时间点通常不可控，对故障的发生过程往往缺乏相应的监控和记录，故障根因分析较为困难。

主动演练：通过应急演练、生产场景全链路压测等方式主动探究系统的稳定性保障体系有助于更加全面地完善系统稳定性体系。一方面，主动演练有助于提前发现潜在的问题，减少突发情况造成的经济损失，而且主动演练过程较为可控，故障的应对过程有充分的时间，

这有利于对故障的根本原因进行定位并制定长期的解决方案。另一方面，有组织的主动演练能够在相对安全的环境下很好地锻炼各方对于突发情况的应对能力，为稳定性保障人员提供更多的锻炼机会。

三、三个关键要素：人员、管理、技术

稳定性是一个整体层面的属性，因此人员、管理、技术三个方面中的任意一项存在短板都会造成稳定性隐患，稳定性保障工作需综合考虑这三个要素。

（一）要素一：人员

根据中国信通院对信息系统故障根因的统计，人为因素导致系统故障的占比超过 90%，包括操作失误、架构设计问题、系统实现漏洞等等。稳定性保障工作即需要参与系统建设和运维的所有人员共同参与，也需要一些负责稳定性保障的专职人员。

整体层面须提升人员的稳定性保障意识：参与系统生命周期各个环节的人员都可能因为人为的失误引入稳定性隐患，这就要求各方都能对自身工作所可能带来的风险进行主动识别，主要可以通过三点予以提升。其一是消除侥幸心理，很多时候风险会被意识到，但由于触发概率看似较低而不易引起重视。随着系统规模、用户群体、系统研发运维团队成员人数的不断增大，这些隐患的触发概率将大幅提高，这就要求系统相关人员对任何可能产生问题的环节引起重视。其二是加强线上意识，系统的上线意味着任何故障的发生都将对用户群体的使用体验造成影响。一些规模庞大的系统拥有亿级别的用户量，且存在较短的时间内高并发访问的情况，一旦故障发生将对业务造成打

击。这就要求相关人员对线上系统抱有敬畏，并认真对待可能影响用户体验的问题。

专业层面注重实践和经验的积累：对于稳定性保障工作较为重视的机构通常会设置站点可靠性工程师（Site Reliability Engineer, SRE）团队或稳定性保障小组，这就要求相关人员具备过硬的专业素养。不仅需要针对监控告警、故障演练、安全发布、巡检等技术具备一定的实践，也须通过经验的积累加深对系统的了解，从而能够更敏锐地定位故障。稳定性保障专业团队通常会注重稳定性知识库的建设，以便加强集体知识的积累。

（二）要素二：管理

稳定性建设和保障工作需要各方配合，因而稳定性相关工作宜从管理的角度着手。保障工作的实施一般需要开发、测试、运维、产品和业务人员等多方参与，其中开发、测试、运维人员为主导方。参与稳定性保障工作的每个团队都需要在团队目标中背负一部分稳定性指标，作为绩效评估的依据，也需针对每个团队制定相应的规范。表 2 展示了各个职位需要制定的稳定性保障规范和相关指标示例，可供参考。

表 2 各职位稳定性保障规范及相关指标

职位	稳定性保障规范	稳定性相关指标
开发人员	代码质量规范	代码语法合规性，可读性等
	日志规范	日志信息的完备程度、时效性
	项目架构规范	解耦程度、冗余程度、隔离程度、承载能力
测试人员	提测规范	代码评审通过率、代码扫描通过率

	测试规范	测试完成度
运维人员	上线规范	测试通过率、测试覆盖率
	预案规范	预案覆盖程度、已验证预案的比例
	复盘改进规范	故障修复率
产品和业务人员	用户反馈收集规范	反馈响应时间
	活动筹办规范	活动期间系统承载率

来源：中国信息通信研究院，2022 年

（三）要素三：技术

稳定性保障工作的技术方面主要体现在韧性架构的使用和稳定性保障工具平台的使用两个方面。

韧性架构：技术架构的设计需结合系统多样化的场景要求，做好相关设计工作。可依据场景的实际情况有针对性地采用一些韧性架构，下面列举了一些韧性架构和使用案例，以供参考。

表 3 韧性架构及相关案例

韧性架构	描述	架构使用案例
冗余设计	对资源留出安全的余量	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理而不存储状态，方便服务崩溃时业务的迁移	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	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	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	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	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	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

稳定性保障工具：稳定性保障相关工具的建设可以极大地提升保障工作的效率，降低维持系统平稳运行的成本，并减少人为错误操作引发故障的频率。下面列举了一些开源的稳定性保障工具或平台。

表 4 稳定性保障相关开源工具

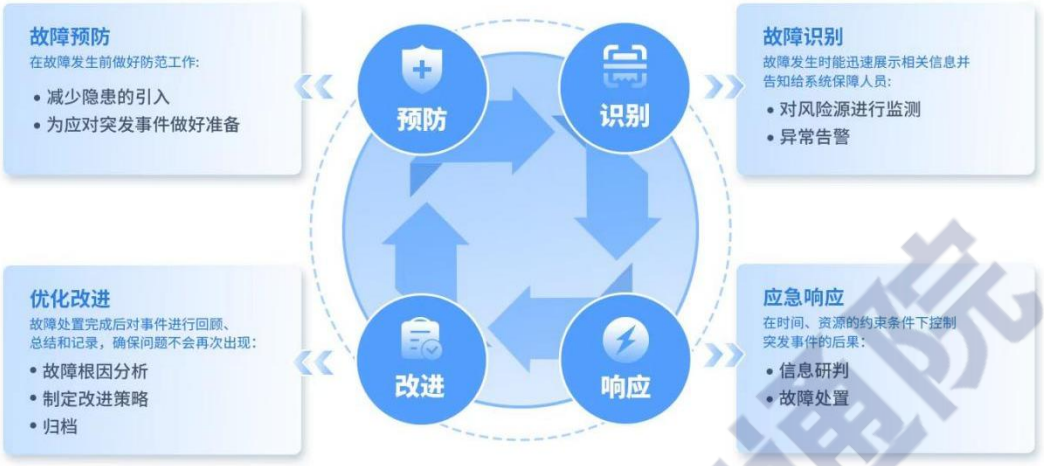
工具类型	工具名称	Github 星数	开源时间
全链路压测工具	Takin	1K	2021 年 6 月
监控告警工具	滴滴夜莺	4.3K	2020 年 3 月
	点评 CAT	16.4K	2016 年 3 月
	zabbix	2.1K	2010 年 7 月
故障演练平台	chaosblade	4.5K	2019 年 4 月
	chaosMonkey	11.9K	2016 年 12 月
	chaos-mesh	4.6k	2019 年 12 月
	chaostoolkit	1.4K	2019 年 2 月
灰度发布平台	kubesphere	8.9K	2018 年 12 月

来源：中国信息通信研究院，2022 年

四、保障系统长期稳定运行的四项核心能力

系统稳定性保障工作的四项核心能力是对系统故障的预防、响应、处置和对不足之处的改进。系统稳定性保障体系的有效性通常是由系统的短板所决定，在时间的维度上也是如此。系统运营方作为一个持续运行的有机体，如果其稳定性保障能力存在短期化的特征，则易导致稳定性保障工作流于表面，给系统的长期稳定运行留下较大隐患。

因此需将这些系统稳定性保障能力融入组织机构的日常活动中，并建立起相应的长效机制。



来源：中国信息通信研究院，2022 年

图 5 保障系统长期稳定运行的四项能力

（一）能力一：故障预防

故障的事前预防是稳定性保障工作的最大头，据中国信通院统计，各方在故障预防上投入的时间在稳定性保障工作总时间中占比超过 80%。提前预防故障的发生将会极大地减少故障可能造成的损失。事前预防能力可分为三部分，一是减少系统在设计、编码、测试、发布变更过程中的隐患引入，二是为应对系统故障等突发事件做好充分准备。

减少隐患的引入：在设计阶段，系统稳定性优化的整体思路是采用必要的韧性设计，防备上游或用户端的异常输入，夯实系统自身的高可用能力，并对下游的返回具备容错性。图 6 对典型的线上业务系统稳定性方面的架设计要点进行拆解，以供参考，通常可采用架构评审的方式减少设计方面的隐患引入。

来源：中国信息通信研究院，2022 年

图 6 线上业务系统稳定性设计要点

在编码阶段，需制定代码规范和日志规范对代码的质量进行要求，可通过代码评审来进行质量把控，发现潜在风险并识别出优秀的工程实践案例。在测试阶段需建立完备的测试规范并对测试覆盖率、单元测试通过率进行要求。在发布和变更阶段需制定相应的规范，可采用变更审批等方式确保流程可控，也可采用灰度发布、蓝绿发布、滚动发布等方式减少风险。

为应对系统故障突发事件做好准备：需要为可能到来的突发事件

做好预案。通常采用以演促防的策略，从而提前发现稳定性保障工作的技术、系统和组织的薄弱环节，提前着手解决问题，完善预案中可能存在的漏洞。演练的方式通常采用全链路压测的手段对系统的最大承载能力进行摸底并锻炼系统扩容能力，或采用故障演练的方式对系统抵御故障的能力进行评估改进。

（二）能力二：故障识别

任何连续稳定运行的生产系统都离不开监控与告警，通过监控来了解系统和业务运行的状态，通过报警来感知监控指标、业务的异常。

对风险源进行监测：完善的监测可以使系统运营方对系统运行情况具备大致的了解，并能在系统异常时及时发现问题。通常采用实时监控、定期的健康巡检、用户端拨测等方式对风险源进行监测。根据被监测的主体可以将监测指标分为不同的层次，如下表所示：

表 5 监控系统的四项黄金指标

层级 / 指标	系统层	应用层	业务层
延迟	系统 I/O 响应时间	读写延迟，虚拟机线程延迟	核心入口响应时间
流量	系统 TCP 流量	数据库或中间件 TPS/QPS	核心入口 TPS 或 QPS
错误	进程中断；TCP、SSH 断开连接	数据库错误；中间件层错误；RPC 线程池满；虚拟机内存使用错误（OOM）	业务成功率；业务数据正确率
状态	CPU 使用率；内存使用率	数据库连接数；RPC 线程池使用率	业务关注指标（业务大盘）

来源：中国信息通信研究院，2022 年

异常告警：在一些指标超出正常范围时，系统会将这些异常状态以报警消息的形式传递出来。告警通常需具备以下维度的考虑：一是告警级别，即当前告警被触发时，问题的严重程度，需针对不同的告警级别设定不同程度的响应方式；二是告警阈值，即一项告警的触发条件，过于迟钝或过于敏感都会引起后续的响应失当，需根据具体场景合理制定；三是通知方式，若为业务指标异常，宜采用较为实时的通知方式（如电话通知等），若为应用、系统层告警，则告警通知中需包含更多关于故障原因的信息，通知方式可考虑钉钉、短信等低干扰方式。

（三）能力三：应急响应

应急响应能力是指在系统故障发生时，应急管理者研判事件信息，并根据故障的性质启动相应的应急预案，开展应急处置工作的能力。

信息研判：应急响应人员需要在极短的时间和较大的心理压力下对现有的信息做出分析，并迅速做出关于故障类别、处置方式的初步判断。信息研判的速度和准确度会在很大程度上影响着系统平均维修时间。快速准确的信息研判取决于以下几点：一是需对事件情况有尽可能详细的掌握，这取决于监控、告警、日志等信息是否完备，二是需对系统架构具备足够的了解，通常参与信息研判的人员需具备丰富的系统相关经验，三是需要和故障相关方进行充分的沟通协调，确保信息充分同步。

故障处置：故障处置能力是指应急响应相关人员在时间、资源的约束条件下，控制系统异常、故障突发事件的后果。故障处置要以快

速恢复系统，降低损失为第一优先级，如故障原因暂时无法定位，通常会将流量切换到备用链路，或采用回滚的方式将系统回溯到稳定的版本。另外，故障处置流程自身的执行质量也会对故障处置结果有重要影响，因此需要对执行质量本身加以监督，所有故障处置流程内容需采用文档或者平台来记录。

（四）能力四：优化改进

日常稳定性保障过程中出现的故障其实也是一个很好的学习机会。需通过应急响应期间的监控数据，分析故障根因，制定后续改进策略，并对故障及应对处置情况进行归档。

故障根因分析：故障根因分析是通过监控、日志、操作记录等信息，结合系统相关知识对故障的原因进行分析总结。其要点是分析故障的深层次原因是什么，是否具有普遍性，以及如何避免再次出现相同的故障。为避免出现信息偏差，建议参与故障处理的相关人员共同参与。

制定改进策略：根据故障根因分析的结果，相关人员应制定系统改进、优化的一系列措施。改进方向有管理方式的变更、规章制度的完善、流程优化、架构调整、扩容、功能优化、代码缺陷修复等内容。改进策略需考虑相应的验收标准，改进结束后需进行验证，确保改进生效。

归档：将这些故障及处置应对措施进行归档，形成故障知识库。这将有助于为后续相同问题的处理提供参考，并对稳定性保障体系的进一步完善提供依据。

五、关键时期稳定性保障的五项重要工作

信息系统在投产后，除了日常的稳定运行外，还会遭遇一些非常规场景，甚至是极端场景，这些极端场景所需要保障措施和平时有较大差别。

表 6 各行业系统关键时期场景案例

行业	系统案例	关键时期场景案例
电子政务	健康码系统、行程卡系统	突发疫情、春运高峰
出行	网约车平台、12306 购票平台	节日期间早晚打车高峰、春运高峰、促销活动
电商	网络电商平台	双十一、年货节、会员日期间大型促销活动
直播	直播平台	直播秒杀、年度直播活动
物流	物流调度平台	双十一、春节期间
金融	银行核心系统	积分兑换活动期间
能源	网上缴费平台	统一缴费日、活动日
零售	零售平台	会员日、广告集中投放营销

来源：中国信息通信研究院，2022 年

在这些关键时期对系统开展重点保障工作是及其必要的，原因有以下三点：一是这些关键时期系统负载远高于日常，且无法根据日常的访问量进行判断；二是这些关键时期的用户行为和日常时期通常存在差别（如秒杀活动中的用户下单将更为集中），更有可能暴露一些日常情况下不易发现的问题；三是关键时期的系统宕机将造成巨大损失，例如，亚马逊 2018 年会员日的宕机事件每小时损失超过 1 亿美元。我们总结了关键时期系统稳定性保障的五项重要工作，供大家参考。

来源：中国信息通信研究院，2022 年

图 7 关键时期稳定性保障的五项重要工作

（一）工作一：团队组织

关键时期的稳定性保障通常需要不同团队间的大量沟通，这就要求机构具备团队间横向拉通机制，对人员进行统一的组织和调度。团队的组织重点需关注以下几点：一是需具备明确的职能分工，以达到故障发生时各团队各司其职、有序应对的效果；二是需建立控制中心（或作战室），将各团队的相关人员集中于同一地点，以便更有效地进行信息同步；三是需公开职责交接的过程，如交接时信息缺乏同步，则会造成工作混乱。重点保障时期角色分工和人员职责可参考下表。

表 7 重点保障时期人员角色安排参考框架

角色	职责
总指挥	负责协调分工以及未分配事务兜底工作，掌握全局概要信息；进行重点决策
事务处理人员	监控系统并处理可能发生事故人员，可根据具体业务场景&系统特性分为多个小团队。团队内部存在域内负责人，与总指挥进行沟通
发言人	对外联络的人员，负责对事务处理内部成员以及外部关注人员信息做周期性信息同步，同时需要实时维护更新事件文档或事故处理记录
规划负责人	负责外部持续性支持工作，例如当大型故障出现，多轮排班

	轮转时，负责组织职责交接记录

来源：中国信息通信研究院，2022 年

（二）工作二：场景及系统情况梳理

在重点保障时期到来前需对业务场景和系统情况进行统一、完整的梳理，这样可以提前发现潜在风险，同时可以使相关人员更好地熟悉情况，提升应急效率。

场景评估：首先需要对即将到来的关键时期进行评估，预先估计可能面临的情况，如高峰期各业务的使用量分布情况、高峰持续时长。其次保障团队需和业务部门紧密合作，对于可能引起系统压力过大的场景，可协同相关部门对业务进行调整。

链路梳理：需要对信息系统进行全链路的梳理，从访问入口开始，按照链路轨迹，逐级分层进行分析，得到信息系统全局画像与核心保障点。现实中，一个系统通常拥有十个以上的访问入口，如果无法覆盖所有链路，可优先选择访问量大的入口作为梳理的起点。可依据链路上各节点的依赖情况、成熟度、历史的故障和修复情况将节点分为不同的风险等级。通常高度依赖其它模块的节点，新发布的节点，历史上出现过故障的节点具备较高的风险等级，须在重点保障时期重点关注。

监控告警梳理：进行系统监控梳理时，可以先从核心、影响系统运行效率的链路开始，按照业务、应用(中间件、JVM、DB)、系统三个层次梳理相应的监控，进行查漏补缺。其次需要检查告警阈值、时间、告警人等配置是否合理。

容量规划：容量规划的本质是追求重点保障时期的风险最小化和成本最小化之间的平衡。需尽量精准测算系统在流量高峰时的负载值，再将负载值根据单点资源负载上限换算成对应容量值，得到最终容量规划模型。业界一般会通过压力测试的方式得到当前系统的容量上限值，然后将其与预估的高峰流量值进行对比，如预估流量峰值高于当前系统容量上限，则需要对系统进行扩容。

（三）工作三：预案准备

在关键时期，仅靠保障人员的临场发挥来来应对线上故障是远远不够。在时间受限的情况下，无法给处理人员留有充足的策略思考空间和试错空间。错误的处理决策，往往会造成更为严重的业务和系统影响，降低用户对系统的信心。因此，要想在故障现场快速而正确的响应，应急保障团队需要依赖提前准备好的应急预案。

应急预案梳理：从执行时机与解决问题属性来划分，预案可分为技术应急预案、技术前置预案、业务应急预案、业务前置预案等四大类，具体描述见表 8。保障团队需结合之前的场景及系统情况的梳理结果，分析出链路中可能出现的风险点，并制定对应的预案：

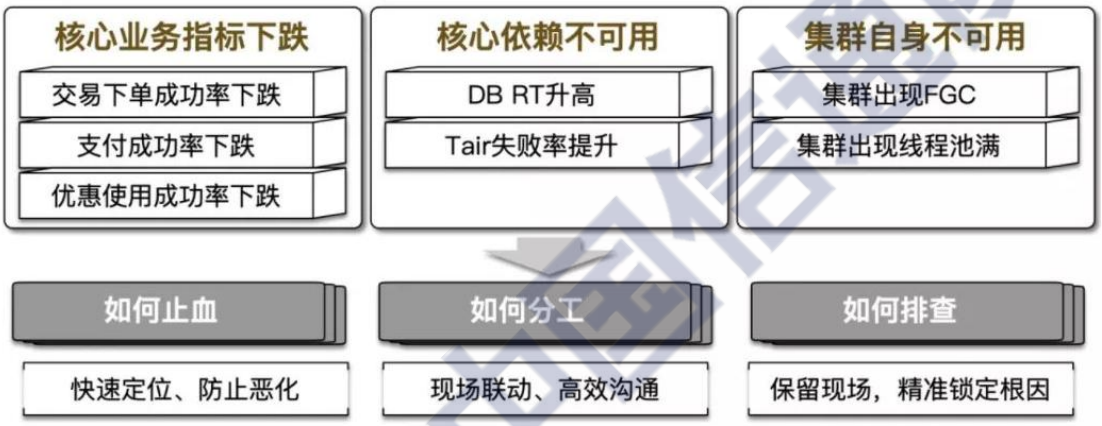
表 8 应急预案分类情况

预案类型	描述
技术应急预案	该类预案用于处理系统链路中，某层次节点不可用的情况，例如技术/业务强依赖、弱稳定性、高风险等节点不可用等异常场景
技术前置预案	该类预案用于平衡整体系统风险与单节点服务可用性，通过熔断等策略保障全局服务可靠。例如弱依赖服务提前降级、与峰值流量时间冲突的离线任务提前暂定等
业务应急预案	该类预案用于应对业务变更等非系统性异常带来的需应急处理问题，例如业务数据错误(数据正确性敏感节点)、务策略调

	整(配合业务应急策略)等
业务前置预案	该类预案用于配和业务全局策略进行的前置服务调整(非系统性需求)

来源：中国信息通信研究院，2022 年

沙盘推演：在重点保障前夕需要确保应急保障团队尽可能熟悉预案内容，对关键预案开展沙盘推演是非常必要的。沙盘推演旨在通过对预案的回顾帮助应急人员提升故障处理能力，着重关注止损策略、分工安排、问题定位等三个方面。



来源：中国信息通信研究院，2022 年

图 8 某线上业务应急预案沙盘推演案例

（四）工作四：事中协同

在重点保障工作期间，需要各方时刻处于备战状态，做好监控和信息同步，如出现线上问题应优先恢复系统的正常运作。

实时监控：关键时期的稳定性保障过程中，各角色应实时关注业务、技术、系统指标监控，以便能够及时掌握最新情况，做出反馈调整。对于核心业务指标需确保监控的时效性，并生成核心业务访问量实时趋势图。对于关键系统指标，如应用接口每秒查询量、响应时间、成功率、CPU 占用率、负载情况等，也需由专人进行相应的监控。

通常可以将这些监控指标进行图形化展示，集成于指挥室监控大屏上，这样将大大减少信息同步的工作量。

信息同步、记录：重点保障工作期间，工作交流群也会出现更多的聊天内容。为了避免漏掉重要的信息，需要采取机制保障客服部门、业务部门和技术部门之间信息的有效传递。

线上问题处理：遇到严重故障时，第一优先级是尽快分析情况，启动相应的预案，尽快将系统恢复正常。需注意的是重点保障时期所面临的情况可能复杂多变，如有必要应结合实际情况对预案进行调整。除此之外，还应做好舆情处理，通知用户等工作，需确保各团队间沟通渠道的畅通。

（五）工作五：事后复盘

当重保结束时，需要对关键时期的整体保障过程进行复盘总结，进行经验的沉淀，并为日常的系统稳定性保障工作方向提供参考。复盘的内容主要包括过程回顾、系统表现回顾、保障工作总结三个部分。

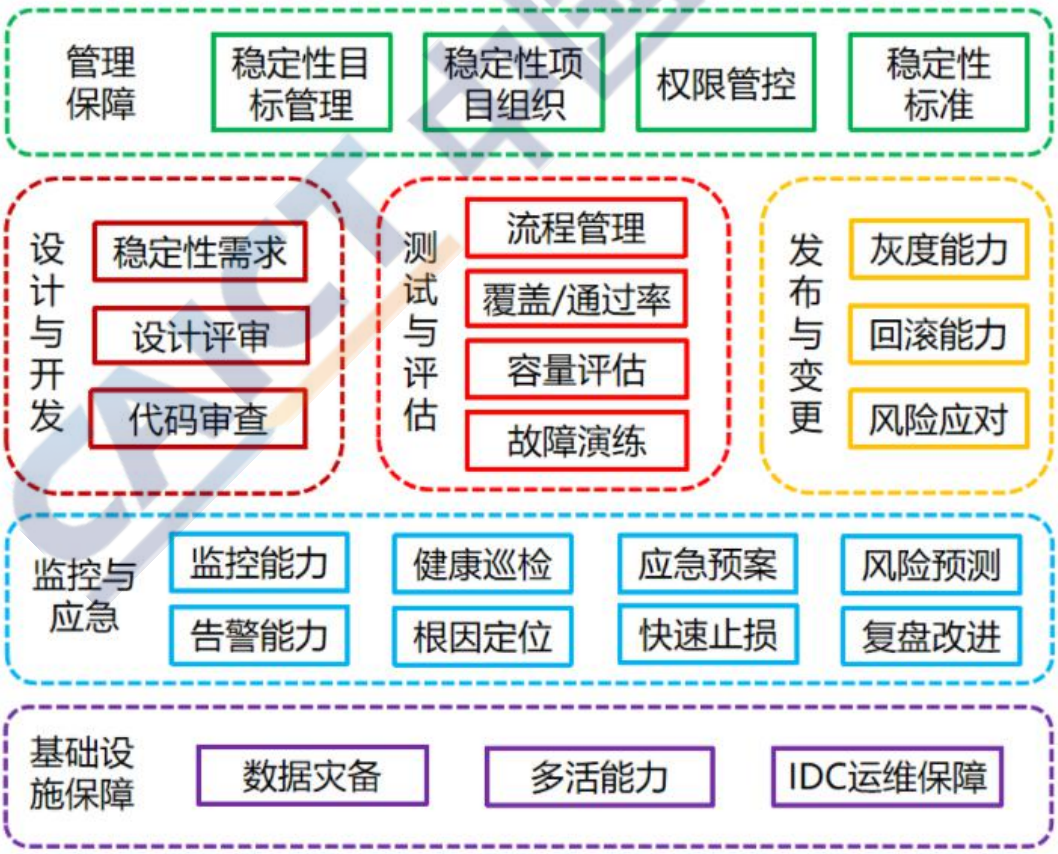
过程回顾：回顾并整理保障工作过程中的各种相关信息，尽可能地把所有情况记录到文档中。

系统表现回顾：对关键时期系统的表现进行回顾和记录，重新评估系统应对类似事件的处理能力。对系统在重点保障期间出现的问题进行整改。

保障工作总结：通过对流量高峰事件进行统计、汇总以及预案执行情况总结，对出现的问题进行分析讨论，不断改进关键时期的稳定性保障预案。

六、稳定性保障工作评估的六个重点方向

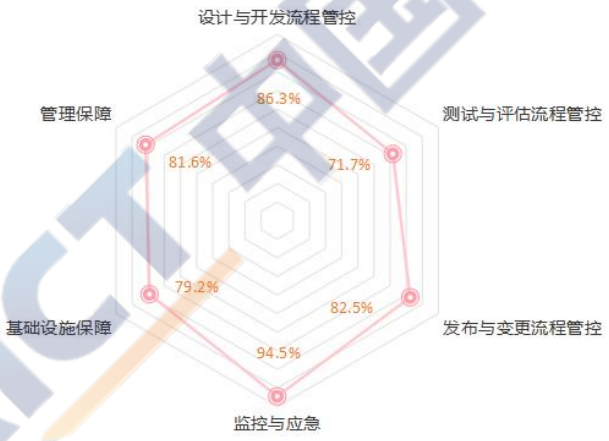
建立信息系统稳定性保障工作的评估体系有助于机构衡量其稳定性保障工作的效果，排查工作中存在的问题，并推动机构内稳定性保障工作的实施。为了形成新阶段信息系统稳定性保障的方法论，建立各个机构可以参考的稳定性保障工作评估体系，提升业界信息系统稳定性保障能力，中国信通院分布式系统稳定性实验室联合 20 多家机构，制定了《分布式系统稳定性保障能力标准》，标准覆盖了信息系统的全生命周期的各类稳定性保障工作。如图 9 所示，标准由基础设施保障、监控与应急、设计与开发、测试与评估、发布与变更、管理保障等 6 大能力域、25 个子能力域构成。



来源：中国信息通信研究院，2022 年

图 9 系统稳定性保障能力标准框架

中国信通院于 2021 年起开展的系统稳定性保障能力评估项目中即采用上述的稳定性评价标准，目前已经完成 10 多个机构信息系统的评估工作，参评系统涵盖电信核心 IT 系统、金融大数据平台、金融微服务平台、金融智能反欺诈平台、政务防疫平台等类型。评估结果汇总见图 10，结果显示各方对监控与应急部分重视度较高，相关工具、制度与实践均较为完备，系统稳定性保障相关的测试及评估工作则相对缺乏。本章将对《分布式系统稳定性保障能力标准》进行展开，详细介绍稳定性保障能力评估的六个重点方向，为各方稳定性保障能力评估工作提供参考。



来源：中国信息通信研究院，2022 年

图 10 信通院首批稳定性保障能力评估平均各项达标百分比

（一）设计与开发流程管控

稳定性要求制定：稳定性要求属于质量要求的一种，需要在产品需求设计阶段制定，并尽可能细分，落实到开发、测试、运维等各角色。

设计评审：系统架构、业务功能方面的评审需要涉及稳定性方面

的考虑，主要体现在两个方面：一是需考察系统是否有针对性的应用韧性架构，如冗余设计、过载保护、去中心化等；二是需要考察系统是否能有效应对异常场景，如在尖峰流量、缓存击穿、网络不稳定等场景下是否能够保持稳定。这些设计也需要具备相应的验证手段。

代码审查：对代码进行自动化或人工审查，并设置代码提交门禁有助于减少低层次稳定性隐患的引入。

（二）测试与评估流程管控

测试流程管理：测试过程需严格、规范、有效，需具备完善的测试报告和测试流程规范。并需要具备一定自动化测试的能力。

测试覆盖率、通过率评估：测试需完整而全面，且需要对开发过程形成有效的反馈。需具备基本的功能性测试如单元测试、集成测试、回归测试等，以及非功能测试如性能测试、安全测试等。对测试覆盖率（行覆盖率、分支覆盖率）、通过率进行统计且需具备相应的要求。

容量评估：需具备针对系统或关键组件的容量评估能力有助于流量高峰时期维持系统的平稳运行。容量评估要求具备对容量进行监控的能力，并可进行全链路（或部分关键链路）的压力测试。

稳定性分析：需具备对系统稳定性或稳定性相关指标进行分析的能力。包括进行稳定性相关测试，对系统模块进行稳定性(健康度)分级，分析组件间的强弱依赖等。

故障演练：故障演练是主动地对故障模拟并对故障进行应对，有助于锻炼系统和系统支撑人员的故障应对能力、容灾能力等，检验历史故障的修复情况，并评估故障应对机制和应对预案是否有效。

（三）发布与变更流程管控

灰度能力：发布时采用渐进的灰度发布策略，有助于减少不稳定版本的影响范围。灰度发布可按照发布对象进行，如按照 IP、机房进行灰度发布，更完善的灰度能力可以按照用户对象的级别进行，如按照用户组别，地域进行灰度发布。灰度发布平台的使用也使得灰度的编排更为容易。

回滚能力：当变更引发系统异常时，快速而有效的恢复到变更前的状态有助于降低故障对生产过程的影响。回滚的效果取决于回滚实施过程以及回滚决策的时效性。决策和实施的自动化将有助于系统的快速恢复。

发布变更风险应对：需对发布变更的风险进行评审。对于高风险的发布需采取一定措施，通过先转移流量等方式使得应用处于非业务处理状态再进行发布，以减少发布风险。

（四）监控与应急

监控能力：对关键的系统资源指标，服务指标进行监控的能力有助于提升系统透明度。监控的完备程度取决于监控的类别与层级是否全面，通常需包括基础架构监控、应用性能监控、网络性能监控与诊断、用户体验监控、日志监控、业务性能监控等，需覆盖链路的不同层级。

告警能力：当系统出现异常时需有效地通知相关人员。完善的告警机制通常具备对告警的分类分级，不同级别或类别的告警具备不同的相应方式、应急水平和处置人员。完善的告警平台可以实现对告警

的自由配置。系统运维方也须定期地对告警进行分析和优化，对冗余的告警进行合并。

健康巡检：健康巡检是指定期对系统更全面的指标进行检查，以便提前于用户发现系统的异常之处。比较完善的自动化巡检系统可模拟用户视角进行自动化拨测，在生产环境中模拟真实用户视角检查系统功能的可用性。

应急预案能力：面对系统可能发生的故障或事件需提前制定标准作业流程，即应急预案。预案需通过演练进行验证，并评估其有效性。

快速止损能力：快速止损的能力取决于支撑后援团队的时效性。需具备应急响应人员，具备合作止损的处理平台以及信息同步平台。对于时效性要求较高的止损任务可具备自动化的故障应对能力，如自动化的限流降级、主备切换等。

风险预测能力：根据现有的稳定性相关指标变化趋势提前预测风险的能力。

故障诊断和根因定位能力：寻找故障、缺陷或隐患产生原因的能力。需具备常见的故障定位标准作业程序和完善的日志查询、服务链路追踪、知识库等辅助支撑系统。

复盘改进能力：修复已发现的稳定性缺陷的能力。包括对系统缺陷开展的修复工作，以及对人员、组织管理相关的失误进行知识库建设和有针对性的培训。

（五）基础设施保障

数据灾备：即对核心数据进行备份和恢复的能力。需对核心数据

进行实时备份，并具备快速容灾切换的能力。需对备份恢复的能力进行周期性地验证。

多活能力：系统所部署的机器或所在地需具备一定的冗余性。包括同机房多活、同城多活和异地多活等不同级别。

IDC 运维：需具备对系统所需的硬件设备进行管理、维护、保障的能力。

（六）管理基础保障

稳定性目标管理：结合业务的实际情况，对业务设置合理的稳定性目标，并跟进目标落实状况。

稳定性项目组织能力：为保障应急工作的顺利进行，需具备快速有效的组织能力。需具备快速响应的稳定性保障团队和横向拉通机制，具备敏捷的项目组织能力。

权限管控：为避免误操作造成的稳定性事故，生产系统的各种权限需满足最小可用原则，且需具备权限申请流程，系统的管理/操作权限需彼此隔离。

稳定性标准：需具备机构内部适用的稳定性相关标准。业务相关的故障需具备故障等级、等级定义等相关说明，稳定性缺陷造成的损失可量化。

七、总结与展望

随着各行各业的现代化演进日趋成熟，各组织机构在运行维护上的投入比例将逐渐增大。组织机构稳定性的提升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加可控。数字经济时代，信息技术已经渗透到生产生活的各个方面，组织机构层面的稳定性在很大程度上取决于信息基础设施能否平稳而连续地提供服务。受政策推动、技术发展、市场需求等多方面驱动，稳定性保障相关产业仍将保持高速增长，其发展趋势如下:

政策导向将加速各方稳定性保障能力建设落地: 2021 年，多款涉及民生的线上系统发生宕机，引起了社会舆论的广泛关注。《关键信息基础设施安全保护条例》也于 2021 年 9 月 1 日正式实施，强调了系统运营者需依照条例、有关法律和国家标准的强制性要求，保障关键信息基础设施安全稳定运行。“速质并重”将是未来我国信息技术产业发展的主基调。

技术发展将推动稳定性保障工作的智能化演进: 随着大数据、人工智能等新兴技术应用于研发运维过程中，数据驱动智能运维的理念逐步被各方所接受。通过对大量监控数据的采集、处理、进行关联性分析，使得对故障的预防、识别和响应能够在一定程度上智能化进行，这将显著提升稳定性保障工作的效率，并降低人为失误发生的概率。

数字化转型将促使稳定性保障产业分工进一步明确: 随着数字化转型进入深水区，大量传统行业如政务、能源、交通开始将各项服务移至线上。相较于互联网等信息技术原生行业，这些领域缺乏稳定性保障相关经验，对稳定性保障服务有较高的需求，这将催生出更多专业提供稳定性保障服务的初创公司，一些信息技术原生企业也逐步推出了专业化的稳定性保障服务。

附录

（一）信息系统稳定性最佳实践案例

1. 能源行业案例——网上国网系统压测与演练工作

国家电网为实现交费、办电、能源服务等业务“一网通办”，建设了网上国网应用程序，注册用户超过 2 亿。只有系统持续稳定运行，才能给广大用户提供更好的用户体验。由于机构组织架构、权责归属环境复杂，存在流程繁复、多环境、多供应商的情况，这些因素增加了系统稳定性保障的难度。针对以上现状，国网开展了以下压测及演练工作:

多环境多轮压测，确保压测不影响业务: 多轮压测共为系统检测出 30 多个问题，针对不同类型的问题均进行了对应的优化工作。为保障压测不对线上业务产生影响，优先在测试环境进行了十几轮压测，确保压测链路、场景、流程无误，最终在生产环境落地实施压测。

应急预案梳理，提前做好响应方案: 为保障系统在面对实际峰值流量时稳定可用，针对各情况梳理预案内容，包括但不限于限流预案、降级预案、熔断预案、隔离预案、防资损预案、容灾恢复预案等系统应急预案。

故障演练，锻炼团队协调作战能力: 协同团队与众多供应商模拟系统故障发生，提前演练沟通，解决跨团队协同难的问题，让团队具备正确处理故障的能力，当系统出现性能瓶颈时快速响应。

针对网上国网核心业务及功能场景进行的生产压测，提前发现并

解决了系统潜在性能问题，有效提升了系统性能。预案梳理与故障演练使得性能保障团队拥有完善的系统故障分析与应对能力，帮助国网建立起系统稳定性保障体系，在每月“缴费日”活动期间从容应对高峰流量，保障用户使用体验。

2. 教育行业案例——学在浙大平台的稳定性保障工作

浙江大学积极响应国家号召，致力于线上教学的发展，推出“学在浙大”平台，受疫情影响学生无法返校上课，系统需要支持全校 8 万师生的顺畅使用，为了保障正常的教学活动不受系统故障影响，浙江大学最终采取了对核心链路进行生产环境全链路压测与监控的解决方案。

采用基于全链路压测的容量规划：浙江大学采用 Takin 压测平台在真实环境中进行测试，并快速定位了问题，为技术人员优化服务器配置提供了量化依据，根据压测数据对各个应用、单个机器、集群服务器进行准确的容量水平评估，给出服务器配置建议，最终节约出 20%的服务器资源，降低了几十万的硬件成本。

采用技术大屏监控，实时掌握动态数据：通过观测云技术大屏监控系统实际运行情况，包括计算资源，机器运行情况；重要业务系统使用的流量、内存、CPU 等状况，当系统出现故障时，便能快速响应并解决。

来源：浙江大学，2022 年

图 11 浙江大学网上浙大监控系统

浙江大学春学期第一堂课按照原教学计划全部顺利开课。学在浙大平台当日总访问量突破 100 万次，在线最高访问量 11 万余次，访客数近 3 万余人。整个疫情期间，教学系统运行稳定。

3. 电信行业案例——浙江移动稳定性保障体系

浙江移动作为首个完成中国信通院系统稳定性保障能力评估的机构，在信息系统的稳定性保障上有丰富的探索和实践经验。在稳定性保障体系的建设上，浙江移动多年来坚持以 SRE 为破局转型，持续提升 SRE 团队的工程创新能力，沉淀出一套多维度、全周期、强实践的稳定性体系。可分为故障抵御、上线发布，交付护航三大体系，贯穿了系统从架构设计到线上治理的整个周期。

故障抵御体系：稳定性最直接的反应就是故障抵御，浙江移动建设 SRE 塔台，实现网元、平台、应用、业务各层数据的融合融通。以

1 分钟发现，5 分钟定位，10 分钟恢复为目标，通过各类运行场景的持续建模，目前已完成 L3 级别的规模化应用，并实现了局部场景下 L4 级自智。

上线发布体系：浙江移动在蓝绿灰度发布的基础上，实现了基于弹性沙箱的灰度发布。融合全容器运行环境的弹性伸缩能力和精准的业务流量控制能力，支撑了从前台应用、中台服务、后台任务的全链路灰度验证能力。同时，可以从地市、工号、手机号等不同维度在生产平面和沙箱平面间做灵活调度，实现真实用户的分级充分测试，并保障系统在发布期间的逃生能力。

交付护航体系：在整个系统周期之始，需要把稳定性需求融入到架构设计中。浙江移动结合纸牌推演和混沌演练，对每项架构点做反脆弱的失效影响分析（FEMA）和优化，确保项目入网前基本完成重点风险的提前布防布控。同时，基于沙箱环境，开展日常的红蓝对抗演练，确保始终对运维各类应急预案的反腐治理能力。

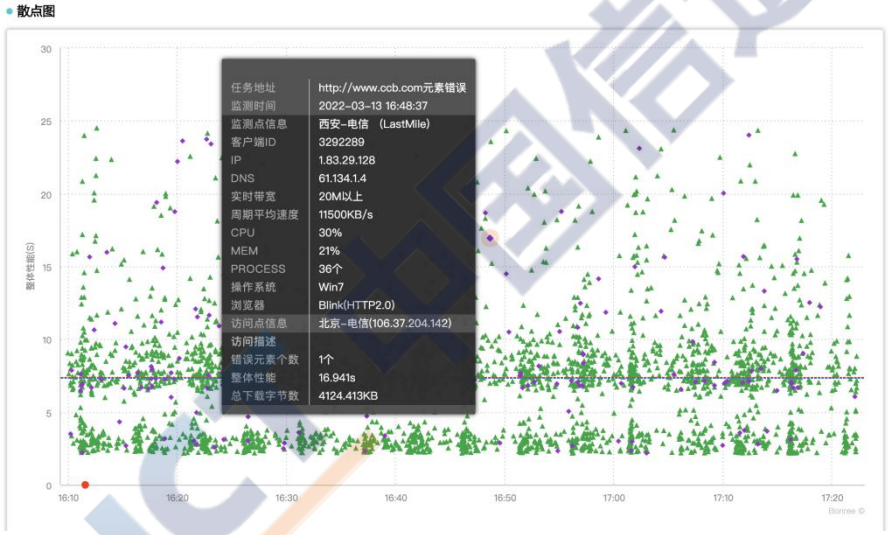
4. 金融行业案例（一）——中国建设银行监控系统

中国建设银行早在 2007 年建行就开始信息化服务体系建设的相关工作。为确保系统的平稳运行，建设银行先后研发了基础监控平台、网络监控平台、业务监控平台等监控系统，但在互联网和移动化的大潮中，发现了这些监控体系依然具备一定的局限性。

根据银行用户体验联合实验室发布的 2017 年《银行用户体验大调研报告》称，银行业离柜业务率（非柜台服务比率）高达 84.31%，手机银行客户满意度仅为 71.8%。而当时建行的主要监控系统都是面

向内部系统的，用户体验问题还会受到运营商网络、CDN 服务质量、终端类型等情况影响。如果无法了解外部用户的真实体验情况，就无法实现故障的快速定位与预警。通过建设互联网监测项目，使用分布在全国各省份主要城市三大运营商的真机用户，对现有电子网银、建行官网、建行云、善融商务、建行大学、主干链路等进行实时高频监测，主动掌握当前业务的系统运行状态。

区域性分析：通过相关监测任务和告警能力，掌握故障/性能的区域性分布，并可通过逐层下钻，实现故障的全面分析与诊断能力。



来源：中国建设银行，2022 年

图 12 中国建设银行散点监控数据分析

散点分析：通过对各个散点数据的统计与分析，实现对单一故障和节点问题的快速诊断，了解相关故障的普遍性及问题的成因。

通过上述的一系列主动拨测手段，并与建行现有的内部监控系统联动，确保了建行网银、建行官网等关键业务系统在全国各运营商用户侧的可用性达到 99.9%以上，并实现故障 5 分钟发现，10 分钟初步定位的能力。

5. 金融行业案例（二）——农业银行反欺诈平台稳定性保障体系的建设

中国农业银行为实现全行欺诈风险一体化管控，建设了企业级智能反欺诈平台，目前已涵盖线上、对公、零售、普惠、信贷等重点领域。为保障业务稳定运行，平台以监控运维、应急恢复、生产复盘、平台高可用建设为抓手，实践出一套包括系统监控、应急处置、容量管理、高可用建设、投产优化以及模型运营等六方面稳定性保障体系。

系统监控：对接行内通用监控工具，通过监控数据采集、存储、告警、可视化等模块，实现了通知告警信息分级分类，根据优先级采用不同的提示手段，有效触达运维监控人员，提前获悉并缓释风险。

应急处置：实现了流量控制、超时控制、熔断控制、服务降级等自动应急手段，并结合应急演练、安全生产复盘、混沌工程实践，不断丰富应急场景，梳理应急处置案例，明确风险巡检、告警处置、应急操作、重启操作等人工应急方案。

容量管理：根据常备测试环境压测结果，分析各个环境的性能瓶颈、容量平衡点，结合上下游系统演练及异常处理所产生的突变流量，评估生产环境的容量平衡点，据此形成巡检指标，及早发现平台容量使用情况、增长趋势，及时进行生产扩容。

高可用建设：通过各业务条线多租户隔离，保障各业务条线内独立运行，不受其他业务条线干扰。通过两地三中心多活建设，有效防范在异地、同城情况下出现数据中心级故障，保障业务连续性。

投产优化：做好源头管理，严格落实变更投产制度，给系统测试、业务验收测试留下充足的时间，保障测试质量及上线质量，坚决斩断缺陷雪球无限滚动，化事后被动为事前主动。结合平台特点实践 DevOps，实现从需求到上线端到端全流程精细化版本迭代发布机制。

6. 金融行业案例（三）——平安科技在版本变更中的稳定性保障实践

平安科技 IT 运维服务于传统金融行业，需要在极高的业务迭代频率和业务复杂度下，保障业务稳定可靠。下面以平安集团在版本变更过程中使用的 D7 自动化应用部署平台为例介绍平安科技的稳定性保障实践。D7 自动化应用部署平台已应用于平安全集团，每月支持超 50 万次变更，自动化执行率达 98%以上，工具异常率控制在 0.01% 以下，极大的提高了版本变更过程的可靠性，对集团各业务系统的稳定性做出了重大贡献。平安版本变更采用集中发布策略，十几年间积累了大量的经验，对版本变更风险有着成熟的控制手段。

在研发领域：对生产环境、测试环境的应用发布，采用统一的管理流程和实施工具，对于版本内容进行隐式安全扫描、合规扫描等。

在运营领域：对于不用进行代码管控、不用测试验证的的简易变更，设置功能开关流程，通过轻量级变更通道进行处理，提升效率；

在安全领域：对于实施过程所需账号等进行独立配置、统一权限、与应用运行账号完全隔离影响；在账号使用过程中，全程免密码，人员不用申请个人的账号

在版本发布实施阶段：发布前，进行版本评审，运营验收，预

发布、全回归发布、灰度或蓝绿发布等，对代码库以及有关会引起发布风险的各项因素进行提前扫描，提高发布的效率以及降低发布风险率。发布中，进行关键字扫描，主动识别和区分使用轻、重量级变更通道，兼顾传统金融行业的严格变更要求和互联网行业快速迭代特性；并在本地、远程端同时备份以及基线处理技术，有问题能及时发现、通知和处理，保障发布过程的流畅，一键回滚等，降低发布过程中的风险。发布后，进行版本自动同步，形成闭环流程管理；以及进行巡检和版本复核，自动调应用可用性平台接口进行可用性验证，能第一时间了解版本发布后的各项应用指标，进一步保证版本发布前后的系统稳定性。

7. 金融行业案例（四）——中广智投监控体系的建设

中广智投秉承让投资更简单的理念，通过中广资本 APP 给客户提供更便捷更智能的投资工具，APP 的稳定性是头等大事。以往都是通过客户投诉被动进行 APP 稳定性保障，既有潜在客户流失的风险，又影响了品牌形象。中广智投采取如下措施，更快速地掌握系统稳定性动态。

实时监控：采用观测云监控数千万客户的应用程序性能和客户体验，有关任何面向客户的问题的详细数据可以实时自动传达给运维团队。

行为分析：通过追踪用户的行为轨迹，全面分析客户的活跃度，留存情况，使用量，关注功能点等，第一时间发现注册新用户，第一时间发现客户关注点，进行有效的 VIP 用户跟踪运营。

主动性运维：通过全面可追溯的信息采集，把页面性能，网络性能，微服务性能，数据库性能，主机性能等关键信息全链路整合，增加了主动性的运维，以线上纠错的方式提前解决潜在通用问题，另一方面下钻到单个用户，解决不容易重现的问题，并为解决客户问题提供上下文。

8. 金融行业案例（五）——中国工商银行企业级对象存储平台稳定性保障实践

企业级对象存储平台是工商银行自主研发的云存储服务，支撑全行各类业务系统，提供非结构化数据联机高并发的存取服务，向接入应用提供基于数据生命周期的分级、清理功能，并具备同业首家海量数据的磁带备份/恢复能力。平台可根据需要采用多园多活的部署模式，以租户隔离、灰度发布及全链路自检能力支撑了全行上百个应用的接入，支持全年 7*24 小时的不间断服务。平台的稳定运行主要得益于 3 大保障体系：

运维监控平台体系：系统的维护是保障稳定性的关键。对象存储平台对接了监控平台、运维信息获取平台、性能容量管理平台、运维巡检平台及综合管理平台等工具。以运维工具研发实现运维操作自动化，以运维信息资产化实现运维信息共享，以数据驱动运维实现运维分析智能化，以运维管控相结合提升运维管理水平。结合 PC 端和移动端，连接开发、测试、生产多个场景，形成全流程运维保障服务。

故障对抗演练体系：为进一步加强设计测试阶段稳定性的融入，在功能和非功能两方面发力。在功能层面，基于对象存储平台核心业

务，持续保鲜场景清单，通过自动化部署流程，在版本交付及投产演练等关键节点开展自动化巡检，确保核心业务的各个场景分支得到全面覆盖，保障业务的可用性。在非功能层面，一是例行化模拟重点接入应用在高峰期的交易状态，使用分布式性能测试平台对系统开展压力及疲劳测试，确保系统对高负载业务场景的稳定应对能力；二是从节点级、集群级、园区级等不同程度的故障场景出发，并借助工行为全行各应用构建混沌测试实验室，注入 CPU 满载、磁盘 IO 冲高、内存高占用、网络延迟堵塞、数据库宕库、容器崩溃等各类典型故障，验证在特殊场景下平台对外服务的高可用能力，以及应急预案在对应情况下的完备性和有效性，及时发现系统架构或应用架构的风险点，进一步提高应用系统的弹性。

全流程线上构建发布体系：软件上线发布平台为行内自主研发，以交付标准化流程为基础，整合持续集成流水线构建及部署、自动化测试、标准化交付流程、应用级投产部署流水线等功能，支持多级别的灰度发布，保障了对象存储平台在每次系统发布期间的系统稳定程度。

通过整个团队的共同努力协作，目前日交易峰值达 2 万笔/每秒，日交易量均值超 4 亿笔，交易成功率稳定在 99.99%，数据可靠性达到 99.9999%，系统故障时 RTO≤10 分钟，RPO≤30 秒。平台的稳定运行，保障了金融业务的有序、平稳的开展。

9. 互联网行业案例（一）——蚂蚁集团稳定性保障体系

蚂蚁集团主要以支付宝客户端提供支付、基金、花呗业务，服

务十亿级用户，业务场景复杂度高，同时涉及金融相关业务，因此对稳定性要求极高。伴随着业务的多年发展，蚂蚁集团逐渐建立稳定性保障方面的问题解决方案和风险防控体系 TRaaS（Technological Risk-defense as a Service）。TRAAS 关注整个研发运维过程可能产生的稳定性风险，从流程制度、文化宣导、技术方案、平台体系多个方面提供稳定性风险防控方案。

统一变更收口，智能变更风险检测：变更核心拦截用户所有发起的人工变更，对接变更统一决策，针对每次变更构建智能化的防控微服务来检测全量风险，实现系统化的三板斧（可监控/可灰度/可回滚）要求，实现智能分批监控、错误码检测、跨链路检测、变更资损检测、变更窗口检测等多种通用防御微服务。

基于 chatops 故障管理，精细化应急定位辅助：基于钉钉群机器人进行顶层入口信息快速汇聚（chatops），在检测故障后在钉钉群自动提示、展示故障信息、展示辅助定位信息、组织相关人员进行相应的应急处理及善后。同时，基于云原生 sidecar 能力，构建了业务调用过程的根因错误、变更信息、业务信息的实时定位图数据，在故障时提供定位辅助信息，以便于相关人员快速定位问题，降低故障影响。

智能资源容量调度，实现稳定性和成本最优平衡：弹性容量基于技术风险防控体系+云原生统一资源调度+数据智能，建设了适合蚂蚁的全局在线资源利用率无风险精确管理和全局容量异常自适应体系，实现了多阶段伸缩、预测式伸缩和容量异常识别和自愈、云原生分时调多种核心技术，支持蚂蚁数千个关键系统，实现 70%的无人值

守率，兼顾稳定性（扩容）和成本最优化（缩容）的综合最优智能决策。

全栈式智能监控，让系统运行更透明，使运维更智能化：蚂蚁智能监控平台自主研发了整套采集、计算、存储和智能化监控产品，建设覆盖了从 IaaS、PaaS 到数据库、中间件、应用以及业务的全栈式、体系化的高可用监控能力，每秒钟监控处理百亿行日志。并且结合蚂蚁高可用场景下的风险识别、应急定位、变更防御等场景沉淀了基于 SQL 的大规模运维数据分析、多场景自适应风险识别、基于多源信息的监控项挖掘、多维下钻分析、关联信息推荐等多项核心 AIOps 能力。

红蓝攻防验证防御能力，527、1218 风险文化宣导：为了提升业务研发运维人员对稳定性方面的意识和经验，同时验证各风险防御平台能力，蚂蚁建设蓝军团队构建攻击能力并进行攻击演练，并在每年的 5 月 27 日和 12 月 18 日举行集中式的稳定性攻防对抗演习。

10. 互联网行业案例（二）——虎牙直播通过业务黄金指标实现稳定性感知和智能定位

传统方式发现问题依靠微服务和基础监控，存在一些不足：告警准确性、有效性得不到信任，会出现用户先于技术人员发现问题，故障根因定位慢的现象。虎牙 SRE 提出了基于业务黄金指标的稳定性感知方法，并实现初因推荐、根因定位、稳定性度量等能力。

实现方法：为了使监控指标聚焦主要业务服务，黄金指标的获取采用用户侧上报的方式，能代表用户真实体验；以黄金指标为监测点

建立端到端链路的可观测大盘。虎牙直播常用关键指标如主播开播、观众进直播间、打赏、订阅等。观测大盘包括全局指标和多维度的细分指标，比如按端、地区、CDN 厂商、平台、用户类型等维度，可进行多维分组对照分析；大盘还包括支撑该指标的后台架构相关的微服务监控、基础监控、网络监控、中间件监控、云服务等，并支持下钻分析。

应用场景及效果：基于黄金指标的监控体系被用于感知业务稳定性，黄金指标抖动必然是业务出现了问题，能代表用户真实质量。感知能力强，无漏告，告警数量少，日均 5 条左右，业务覆盖 100%。黄金指标按时间精度聚合后成为周期性稳定性的量化评估工具，故障发现、故障评审复盘都以黄金指标作为依据。通过黄金指标结合 AIOps 算法分析，提供无阈值告警和阈值推荐能力，算法加上关联全链路的软件元数据实现良好的故障定界、初因推荐、根因定位能力。

（二）国内系统稳定性保障相关服务商

随着各方对稳定性相关工作越来越重视，提供系统稳定性保障服务及相关平台建设的初创企业纷纷涌现。于此同时，一些传统的运维服务商也开始进入稳定性保障服务领域。下表汇总了一些稳定性相关服务提供商，供各方参考。

表 9 国内稳定性服务供应商

企业提供服务类型	代表企业	旗下产品或服务平台
监控告警	蚂蚁数字科技	TRaaS-RMS
	优维科技	hyperinsight 超融合监控
	驻云科技	观测云

	嘉为蓝鲸	监控中心(UMC)、告警中心(UAC)
	新炬网络	ZnAiops 统一监控告警平台
	云智慧	云监控
全链路压测	阿里云	PTS
	腾讯云	WeTest
	数列科技	Takin
故障演练	PingCAP	ChaosMesh
	阿里云	AHAS 故障演练平台
	蚂蚁数字科技	TRaaS-HAS
	新炬网络	ZnAiops 混沌演练平台
灰度发布	阿里云	分流灰度发布平台
	腾讯云	Serverless 灰度发布平台
	蚂蚁数字科技	CAFE
	华为云	分流灰度发布平台
拨测	博睿宏远	bonree APM
	听云	听云 APM

来源：中国信息通信研究院分布式系统稳定性实验室，2022 年

中国信息通信研究院分布式系统稳定性实验室

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