CAICT 中国信通院

从近年运行事故看企业
如何构建云+应用运行安全策略

季可航

中国信息通信研究院 云计算与大数据研究所 云计算部工程师

CAICT 中国信通院

01. 背景

02. 运行事故回顾及分析

03.云+应用运行安全策略构建思路

应用&云服务运行事故影响巨大，亟需提升运行安全重视程度

CAICT 中国信通院

◆ **应用&云服务已成为支撑数字经济和社会运作的基石，一旦发生大范围故障，不仅带来直接的生产力下降和经济损失，还可能损害其长期积累的品牌信誉与公众形象，甚至带来重大合规风险和监管惩处，需要给予充分重视与投入。**

生产力下降

服务中断将直接对企业的运营产生负面影响，还会波及到关联客户、供应链乃至整个行业，造成连锁反应。

2023年，由于服务多次宕机，星展银行被暂停非必要业务6个月，这种惩罚直接限制了其业务范围，减少了收入来源，同时也对客户造成了不便，影响了银行的生产力。

经济损失

服务中断会导致直接的收入损失，包括交易无法完成以及由于服务中断而产生的赔偿。间接损失则来自于客户信任度下降，可能导致长期客户流失。

根据Gartner的研究，IT停机平均每分钟的成本在2014年就达到了5600美元，而具体到大型企业，这个数字可能更高。

社会舆情

近年来，云服务宕机事件频发，大平台宕机极易攀升至热搜榜单，激发广泛的社会舆论反响，对涉事企业公众形象造成深刻且持久的负面影响

7月19日微软宕机事故已有形成度词条

“微软”搜索指数上升4倍

合规风险

对于数据可靠性和业务连续性严格要求的行业，容易带来重大合规风险和监管惩处

由于某证券公司APP多次宕机，监管责令改正

CAICT 中国信通院

01. 背景

02. 运行事故回顾及分析

03.云+应用运行安全策略构建思路

案例1：微软系统全球故障

CAICT 中国信通院

故障概况

北京时间2024年7月19日，一场由第三方安全解决方案提供商Crowdstrike的**组件更新触发的兼容性冲突**，导致了全球范围内的Windows系统用户遭遇了严重的“蓝屏死机”现象。这一突发事件不仅影响了个人用户的日常使用，更引发了连锁反应，使依赖微软Azure云服务的重要机构业务被迫中断。

故障原因

第三方供应商

1.代码质量管理能力薄弱

2.变更验证机制存在不足

应用集成方/平台服务方（微软）

1.对第三方应用的运行安全管控不足：未对第三方变更执行二次检验

2.缺少有效恢复手段：未快速回滚至稳定版本。

3.集成与解耦的平衡：集成第三方组件能丰富功能和提升效率，但过度依赖可能导致系统整体的脆弱性增加

云上应用方（航空公司、银行等）

1.系统架构单一性风险：过度依赖单一操作系统（如Windows），放大了系统级风险

2.应急响应缺陷：缺乏即时的备选方案与快速恢复流程

3.应用自恢复能力缺失：应用程序过分依赖外部供应商或基础设施的恢复，忽视了内置自恢复机制的设计与实施

优化建议

加强软件更新验证能力

- **全面变更验证**：确保变更验证覆盖日常操作及极端故障情形，强化对Windows、Linux、Unix等主流操作系统的兼容性测试。
- **强化变更管理**：建立详尽的变更应对计划，一旦变更引发异常，可通过版本回滚等手段实现服务的迅速恢复。

加强第三方组件治理能力

- **第三方更新后验证流程：**评估第三方更新后与内部系统之间的兼容性。
- **第三方故障应对预案：**开发详细的应急预案，具备应对第三方故障的独立逃生能力。

加强服务容灾能力

- **构建多源容灾架构**：采用多供应商、多地域的分布式部署策略，建立多源容灾架构。
- **强化应急响应与冗余机制**：强化应急响应，包括但不限于实时监控、故障自动切换、快速回退方案等，运用容灾冗余技术如热备、冷备和温备等手段，缩短故障恢复时间。

案例2：ChatGPT等大模型服务多次服务中断

CAICT 中国信通院

故障概况

北京时间2024年6月4日下午至6月18日凌晨，ChatGPT、Gemini、Claude、Perplexity等大模型搜索引擎发生服务中断故障，影响全球近5亿用户，造成极大社会影响。

故障原因

- **数据库配置不当：**ChatGPT数据库因容量不足且配置不合理，在遇到突发流量后发生数据库服务阻塞不可用，从而导致ChatGPT服务宕机。
- **优化方案有缺陷：**6月5日ChatGPT对数据进行优化，但方案存在问题，数据库缺陷仍未解决，导致6月18日再次发生数据库故障。
- **应急处置流程缺少验证：**ChatGPT数据库发生异常后，尝试将流量切换到其他备份数据库，但均恢复失败，主数据库仍然无法访问

优化建议

强化流量冲击应对能力

企业需要建立资源容量的实时**监测评估**与**容量伸缩**机制。强化网络流量的实时监控能力，面对大流量冲击时，能迅速采取**资源扩容**、**服务降级**等有效措施，保障核心服务资源充足。

补全应急处置能力

企业需完善应急管理制度，明确**应急操作流程**，完善**应急预案**，覆盖各类典型故障和极端场景，强化应急预案的**科学制定**、**实战演练**与**适时更新**，增强对突发事件的应对能力。

提升故障优化有效性

企业需提高**故障优化有效性**，故障优化方案需经过**测试**与**验证**。在故障优化完成后，需对系统进行故障场景下的模拟和验证，验证优化方案对于此类场景的有效性。

案例3：必应搜索API全球故障

CAICT 中国信通院

故障概况

北京时间2024年5月23日14时，微软旗下搜索引擎必应（Bing）发生全球性故障，同时也引发其他依赖必应应用程序接口（Bing API）的搜索引擎和服务故障。例如**国外DuckDuckGo、Ecosia、Qwant等搜索引擎无法生成搜索结果；微软人工智能辅助工具“领航员（Copilot）”出现持续加载状态；美国AI聊天助手ChatGPT提示错误反馈。**

故障原因

- **解耦合架构缺失：**当必应引擎发生故障后，企业未能识别故障并进行切换或者降级，导致故障暴露，造成客户使用受到影响。
- **应急处置能力缺失：**本次故障历时长，社会影响大，体现必应缺少故障应急处置能力，未能优先抢通业务降低故障影响。

优化建议

服务韧性架构优化

企业需要持续优化系统架构，保证系统面对故障时的稳定性和服务韧性。本次故障暴露出的架构缺失有以下几点：

- **解耦合架构缺失：**企业需降低系统各服务间耦合程度，避免因为单一服务异常导致系统不可用。
- **容灾架构缺失：**当外部服务不可用时，需具备容灾切换能力，通过将流量切换至其他供应商或者内部服务实现业务恢复。
- **熔断架构缺失：**当确认服务异常后，系统需具备服务降级熔断能力，降低故障影响面。

完善云服务故障应急管理机制

企业需持续完善云服务故障应急管理机制，加强云服务系统运行安全能力。

- **补全应急预案，**持续补充故障应急预案，使其能覆盖所有常见以及重大故障场景，保证故障发生后能尽快恢复业务，降低故障影响。
- **开展应急演练，**企业需依据应急预案定期开展故障应急演练，验证应急预案的可用性和时效性，并提升人员面对故障的应急处理能力，提高故障恢复速度。

案例4：谷歌云配置错误导致UniSuper账户删除

故障概况

2024年5月2日至9日，**谷歌云（Google Cloud）**发生互联网史上最大的云服务运行事故，因“意料之外的配置错误”导致**澳大利亚养老基金管理公司（UniSuper）的谷歌云账户遭到删除，超过60万名基金成员近7日无法访问其退休金账户，涉及金额高达约1240亿澳元**，造成极大社会影响。最终，通过UniSuper的多元异地备份策略，利用除谷歌云外的另一云服务商备份数据，实现了极端场景下的数据重建与恢复，使得业务服务恢复正常。

故障原因

- **配置错误：**由于云服务平台关键配置错误，导致UniSuper账户被删除、相关异地备份数据被清空。
- **变更计划不充分：**未能准确识别本次配置修改的风险点，且未能在配置修改前进行数据的预备份，导致数据彻底丢失。

优化建议

云服务商

云服务商需要建立完善的云服务运行**安全管理制度**，定期开展整治，检验企业在高风险操作、极端场景下应对能力。

- **强化变更风险识别能力：**确保系统变更时经过充分测试，降低因配置错误导致的系统中断风险。
- **建立分级分类控制：**对**业务、操作、事故**等进行分级分类，强化高风险操作的管控、校验和审批。
- **优化数据恢复策略：**企业需强化备份恢复预案的完备性和有效性，保证极端情况下的数据可恢复性。

云上应用方

云上应用方应建设**多元异地备份策略**，强化备份恢复预案的完备性和有效性，针对数据完整性受损的极端场景，设置分级恢复预案，组织定期演练保障预案有效性。

- **多副本备份：**保证单一副本数据丢失不会影响数据重建与恢复。
- **多地备份：**保证单资源池数据丢失不会影响数据重建与恢复。
- **多云备份：**保证极端场景下单运营商数据丢失不会影响数据重建与恢复。

CAICT 中国信通院

01. 背景

02. 运行事故回顾及分析

03.云+应用运行安全策略构建思路

云+应用运行安全策略：技管结合，标准引领，立体提升运行安全水平

CAICT 中国信通院

管理

- 强化权限管理及流程管控
- 健全故障演练机制
- 强化应急处置能力
- 强化分级分类管理

设计

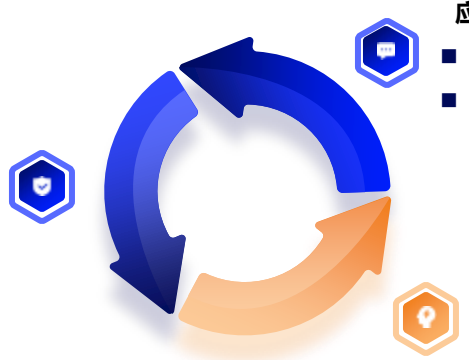
01 增强完整性

02 强化执行力度

落地

- 强化制度执行落实保障
- 健全责任追究与激励机制
- 强化底线思维极限思维

技术



日常运维
- 容量管理
- 监控巡检

应急演练
- 演练场景
- 演练周期

架构优化
- 解耦设计
- 容灾冗余设计
- 限流熔断设计
- 可观测设计

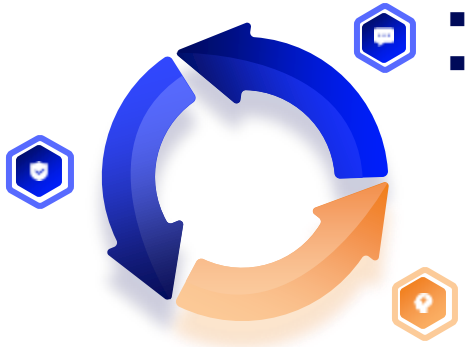
主动优化

- **日常运维途中发现系统架构缺陷**，发起系统架构层面优化行动，避免此类故障重复发生。
- **系统架构优化完成后进行应急演练**，通过演练的方式验证系统架构优化的有效性，同步更新相应的应急处理预案。
- **经过演练验证的应急预案加入日常运维**，依据应急预案处理每日故障，验证应急预案的可靠性和全面性。

被动应对

- **新的需求通过变更方式落地**，从业务、运维、产品方提出的对系统的新需求，依据变更手段转换成系统能力。
- **新的变更为系统引入未知故障**，变更为稳定的系统增加新的变量，从而引入新的风险，可能导致新故障的发生。
- **故障发生触发应急处置**，系统维护人员通过规范的流程和优秀的工具，降低故障对系统稳定运行的影响。

需求来源
- 业务需求
- 运维需求
- 产品需求



应急管理
- 应急响应
- 故障发现
- 故障定位
- 故障处置
- 故障优化

变更管理
- 变更准备
- 变更执行
- 变更验证
- 变更回滚
- 工具化能力建设

标准

云服务运行安全标准体系

制度规范　技术工具　系统架构　人员能力

云+应用运行安全管理策略：增强制度完整性，强化制度执行力度

CAICT 中国信通院

增强完整性

01 强化权限管理及流程管控

完善变更操作管理机制
加强高风险操作技术防护

02 健全故障演练机制

科学制定故障演练方案
定期开展故障演练活动
加强资源和技术投入，提高故障真实性与复杂性

03 强化应急处置能力

定期更新应急预案。
构建一体化应急响应流程
加强跨部门、跨业务联动应急演练

04 强化分级分类管理

实行业务、操作、事故分级分类管理
强化分级分类开展代码质量控制

运行安全管理能力

强化执行力度

01 强化制度执行落实保障

坚决杜绝“纸上谈兵”现象，提高制度执行的严肃性和有效性。

02 健全责任追究与激励机制

强化企业员工自觉遵守相关制度的内生动力。

03 强化底线思维极限思维

强化培训力度，培养员工的底线思维和合规意识。

云+应用运行安全技术策略（1/2）：主动优化，强化能力建设

CAICT 中国信通院

主动开展架构优化、故障演练，构建主动优化内循环

1

日常维护

- 推进**容量管理**，高效管理资源的分配和利用
- 提高**告警巡检覆**盖率和准确率，提前发现系统风险隐患

■ 云服务商

- 在容量冗余和资源高效利用之间取得平衡
- 打通上下游监控通道，全链路监控服务运行状态。

■ 云上应用方

- 建设多环境、多规格、多平台资源储备
- 建立云+应用一体化监控体系，保障服务稳定运行

日常运维发现系统架构薄弱点，发起架构优化

2

架构优化

- **解耦合设计**：系统与各服务模块以及外部服务需进行解耦，避免单一模块或外部服务异常导致系统服务不可用。
- **容灾冗余设计**：增加一套以上完成相同功能的系统，保证当该部分出现故障时，系统仍能正常工作。
- **限流熔断设计**：通过限流、熔断、负载均衡等技术手段，避免因网络攻击、大流量冲击导致的系统服务过载。
- **可观测性设计**：系统需具备采集，上传，分析数据的能力，实现系统状态和行为的可量化以及可分析性。

■ 云服务商

- 具备本地多多活和异地多活架构
- 具备本地多机房架构，且可以提供给云上应用方

■ 云上应用方

- 多云冗余容灾架构，避免单一云服务商故障影响业务运行。
- 云上服务副本具备跨机房、跨资源池架构设计

通过故障演练验证架构优化结果和应急预案。

3

故障演练

- **演练场景**包含常见故障场景以及重大故障场景
- 进行基础设施故障、云平台故障、云上应用故障**全链路故障演练**
- 提高演练频率，故障**演练常驻**

■ 云服务商

- 协助云上应用进行云平台故障演练
- 定期开展基础设施故障应急容灾演练

■ 云上应用方

- 定期开展单一云服务商故障应急容灾方案

在日常维护过程中检验应急预案全面性

云+应用运行安全技术策略（2/2）：被动应对，提升平台健壮性

CAICT 中国信通院

被动开展变更管理、应急管理，强化被动应对外循环

1

需求来源

- **业务需求**
- **运维需求**
- **产品需求**

通过变更落地需求

2

变更管理

- **变更准备：**提升变更风险预估能力，变更方案需经过测试和验证
- **变更执行：**严格遵守变更管理规范，强化流程管理，确保高风险操作有明确的审批链条和权限限制
- **变更验证：**变更后需进行全面系统验证，提前干预未知故障
- **变更回滚：**具备全面有效回滚方案，避免变更引入故障
- **工具化能力建设：**持续建设变更工具化能力，避免人为操作引发故障

■ 云服务商

- 云平台变更需做好客户隔离
- 平台需具备数据迁移能力，当云平台变更异常时快速迁移客户数据

■ 云上应用方

- 变更方案结合云平台数据恢复和迁移能力，提高应对变更问题的恢复能力

变更引入未知隐患触发故障

由于故障发生客户和系统提出新需求

3

应急管理

- **应急响应：**明确响应角色，保证应急响应要求下放到个人
- **故障发现：**监控覆盖全链路，实现端到端的故障感知
- **故障定位：**具备全方面诊断工具，实现快速问题原因定位
- **故障处置：**具备工具化应急处置能力，如流量切换、一键隔离等
- **故障优化：**优化方案需经过故障场景二次验证有效性

■ 云服务商

- 云平台故障时需及时通知受影响客户及下册业务，并协助进行故障逃生。
- 告警覆盖上层基础设施，且具备基础设施故障应急恢复能力

■ 云上应用方

- 具备对上层基础设施和云平台故障的响应和处置预案
- 通过多个供应商实施冗余，以确保恢复措施的多样性和可靠性。

中国信通院 云服务运行安全标准体系

CAICT 中国信通院

运行安全维度

管理：组织架构 | 制度规范 | 人员能力

技术：架构设计 | 技术工具

AI云运行安全

- 《基于人工智能的云计算运维能力成熟度模型》
- 《基于大模型的可观测性能力成熟度模型》
- 《面向大规模智算服务集群的稳定运行能力要求》

运行安全行业专项

通用行业

- 《服务韧性工程（SRE）能力要求》
- 《“云+应用”一体化运维管理能力要求》
- 《一云多芯建设成效成熟度模型》
- 《算力稳定安全运行能力要求 第1部分：算力中心域》
- 《算力稳定安全运行能力要求 第2部分：分布式算力域》
- 《分布式系统稳定性度量模型》

金融行业

- 《证券期货行业重要系统稳定性保障模型》
- 《金融业务连续性成熟度模型》
- 《金融分布式系统风险管理成熟度模型》

通信行业

- 《通信行业稳定性保障模型 第1部分：云平台域》
- 《通信行业稳定性保障模型 第2部分：云上业务域》

运行安全技术能力

1. 1、架构容灾
   - 《应用多活架构能力要求》
   - 《应用多活成熟度模型》
2. 2、研发测试
   - 《全链路压测平台能力要求》
   - 《模糊测试平台能力要求》
3. 3、变更管控
   - 《变更管控能力成熟度模型》
4. 4、监控、日志与度量
   - 《可观测性平台能力要求》
   - 《可观测性能力建设成熟度模型》
   - 《根因分析平台能力要求》
5. 5、平台交付
   - 《云平台交付能力要求》
6. 6、容量管理
   - 《智能容量管理技术成熟度》
7. 7、中间件管理
   - 《API全生命周期管理能力》
   - 《分布式消息队列能力要求》
   - 《分布式事务服务能力要求》
8. 8、可靠可控
   - 《一云多芯稳定性度量模型》
9. 9、故障演练
   - 《混沌工程平台能力要求》
   - 《混沌工程成熟度模型》

全流程运行安全

可信云服务稳定性专项评估

- 消息队列稳定性评估标准
- 云数据库稳定性评估标准
- 容器集群稳定性评估标准
- 云缓存服务稳定性评估标准
- GPU集群稳定性评估标准
- 融合云DNS服务稳定性评估标准
- 云存储稳定性评估标准
- 云主机稳定性评估标准
- 云网络服务稳定性评估标准
- 内容分发稳定性评估标准
- GPU云服务稳定性评估标准

CAICT 中国信通院

THANKS!

系统稳定性建设负责人：季可航
邮箱：jikehang@caict.ac.cn