

华为云 确定性运维

第五期

内部发行
免费赠阅

前沿观点

大咖访谈

技术创新

业务提升

运维动态

让运维成为智能世界变革的加速器

浅谈人工智能运维的发展与新机遇

— SRE专委会 (SRE社区) 发起人 刘峰

LLM和Multi-Agent在运维领域的实验探索

— 华为云专家 张曦

乘云启航, 打造面向未来的数字化新零售商业体

— 百联全渠道电子商务有限公司副总经理 王善良

双轮驱动, 构建运维“数循环”——数据驱动运维的方法与实践

— 云智慧副总裁 陆兴海 等

大模型与智能运维的双向奔赴

— 中山大学教授 博士生导师 陈鹏飞

AI大规模训练集群稳定性实战, 加速智能跃迁

— 华为云专家 童琳



主 办	华为云计算技术有限公司SRE部
顾 问	高江海 安 宇 薛 莹
主 编	林华鼎
副主编	汪庆愿
撰 稿	陈鹏飞 黄 晓 晋 彪 康 镇 李国强 刘 峰 刘纯纯 陆兴海 董 琳 王善良 王福强 伍 杰 杨金全 殷阁朕 张 曦 郑 磊 （按姓氏首字母排序）
编 辑	贺丽萍 韩江荟 李超群 吴文辉 张 飞 （按姓氏首字母排序）
联 系	电子专刊、索阅、投稿、建议和意见反馈请联系SRE部 邮箱: snzx1@huawei.com
地 址	东莞市松山湖高新技术产业开发区环湖路9号华为溪流背坡村
邮 编	523830
无担保声明	本资料内容仅供参考, 华为云计算技术有限公司不对本资料所有内容提供任何明示或暗示的保证, 包括但不限于适销性或者使用与某一特定目的的保证, 在法律允许的范围内, 华为云计算技术有限公司在任何情况下都不对使用本资料任何内容而产生的任何特殊的、间接的、继发性的损害进行赔偿, 也不对任何利润、数据、商誉或预期节约的损失进行赔偿。（内部发行, 免费赠阅）
版权归属	版权归华为云计算技术有限公司所有, 保留一切权利。 非经本公司书面许可, 任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部, 并不得以任何形式传播。

AI

目录



前沿观点

浅谈人工智能运维的发展与新机遇

01

LLM 和 Multi-Agent 在运维领域的实验探索

07



大咖访谈

乘云启航，打造面向未来的数字化新零售商业体

15

双轮驱动，构建运维“数循环”-- 数据驱动运维的方法与实践

21

大模型与智能运维的双向奔赴

30

AI 大规模训练集群稳定性实战，加速智能跃迁

37



技术创新

大模型时代下的应用可观测性方案探索与实践

43

云服务过载控制“三部曲”

49



业务提升

政务云架构设计与应急切换能力优化 实现业务安全稳定高质量

53

租户业务架构韧性提升 保障上云用云稳定可靠

57



运维动态

24 年维享会重点活动回顾

61

推荐语

黄强元

瑞幸咖啡
副总裁

随着人工智能技术的不断发展，AI大模型训练集群的稳定性挑战也随之而来。在《AI大规模训练集群稳定性实战，加速智能跃迁》这篇文章中，作者详细阐述如何利用人工智能技术来解决复杂的业务问题，分享AI集群构建方面的完整闭环能力及实战经验，从异常感知、诊断到故障自愈的全过程，为云上系统的稳定可靠提供有效的保障系统，同时介绍如何通过技术手段来解决AI集群的稳定性问题，不仅推动智能技术的跃迁，也为企业的数字化转型提供了有力的支持。这篇文章可以帮助企业结合人工智能技术来提升企业竞争力，是一篇非常有价值的读物。

米鹏辉

华为GTS交付应用开发与
SRE部 部长

随着各行各业数字化转型的不断深入，IT系统的复杂程度也在不断加深。如何在面临各种不确定性因素的挑战下，构筑IT系统确定性运维能力一直是SRE领域的关键难题。

《华为云确定性运维专刊》积累了华为云多年大规模、复杂系统的运维经验，同时也连接了各行各业的运维专家和先进企业，既有“开天窗”的前言运维技术洞察，也有“接地气”的华为云自身以及各行各业的宝贵运维经验分享，从理论到实践，指导产业真正的落地确定性运维理念，是行业从业者不可多得、也不容错过的运维知识财富。期待与行业专家们一起，共同探索、实践、夯实确定性运维之路，为企业数字化转型保驾护航。

曾华山

珠海金山办公软件有限公司
运维总监

近年来，在互联网普及与经济全球化的趋势下，如何在易变、不确定、复杂和模糊的VUCA时代寻求服务的确定性和稳定性，已成为企业服务质量和运维的头号挑战。

『确定性运维』如一剂良方，深入探讨了在不确定环境中，紧跟AI时代技术发展，通过高可用架构设计、应急演练、AIOps等一系列运维策略，确保业务系统的稳定和可靠。本期结合大量实际案例，展示了如何在不确定性中寻求确定性，为企业在智能化转型的道路上保驾护航。

在AI大浪潮时代，确定性运维，未来可期。

“序言

让运维成为智能世界变革的加速器

云计算带来的普惠技术，成为推动数字化转型的重要引擎。越来越多企业将核心业务部署到云端，华为公有云业务保持了高速增长，客户来自千行万业，使用云的能力参差不齐。如何提供安全、稳定、高质量、易用的服务，是我们要面对的一个挑战，这对云底座的“质量”提出极为严苛的要求，需要采取一系列措施来预防和解决。如何有效管好云，持续为业务赋能，是企业关注的重要命题。

安全可信是前提，稳定可靠是基础，资源高效是关键，业务敏捷是核心。

“确定性运维能力”是我们运维最主要的能力，我们要从设计、开发、部署、监控、运维等环节发力进行全面质量管理发力，对故障发生、故障范围、故障恢复、故障影响做到确定性控制，实现最小化、网格化的管理，从流程、意识、质量文化、考核、工具等，帮助企业完成从传统维护团队向IT平台服务的转型，将业务高速发展带来的“不确定性”变成SLO（服务质量目标）的“确定性”，是我们不断夯实的基础能力。

去年9月华为云正式发布维享会（确定性运维经验交流分享会），华为云希望建立这样一个平台链接全球客户，与众多行业大咖一起探讨新技术、碰撞新思想，通过专刊、白皮书和案例集等载体，进行观点和实践经验分享，探讨业务上云后的最佳实践和创新融合，相互碰撞和启发，共同构建安全可靠的“确定性”运维世界。



高江海

华为公有云业务部总裁





人工智能运维的发展与新机遇

背景

由于大模型等新一代人工智能技术的突破和高速发展，为 SRE 和智能运维（AIOps）领域提供了全新的发展机遇和广阔的创新空间，本文从人工智能运维的定义、大模型的发展、SRE 结合 AI 实现人工智能运维以及带来的改变等多个角度展开分析和探讨。



刘 峰

SRE专委会（SRE社区）发起人，中国早期SRE讲师、GitLab金牌讲师，SRE布道师。

人工智能运维的定义

智能运维是将人工智能（AI）应用于 IT 运维的实践

“AIOps”（人工智能运维）这个术语是由Gartner在2016年创造的，最初用于描述一个新兴的行业类别，旨在使用ML（机器学习）来解决“运维超大规模云基础设施”所产生的问题。AIOps 作为一

种实践与前面的 ITOA（IT 运维分析）密切相关。

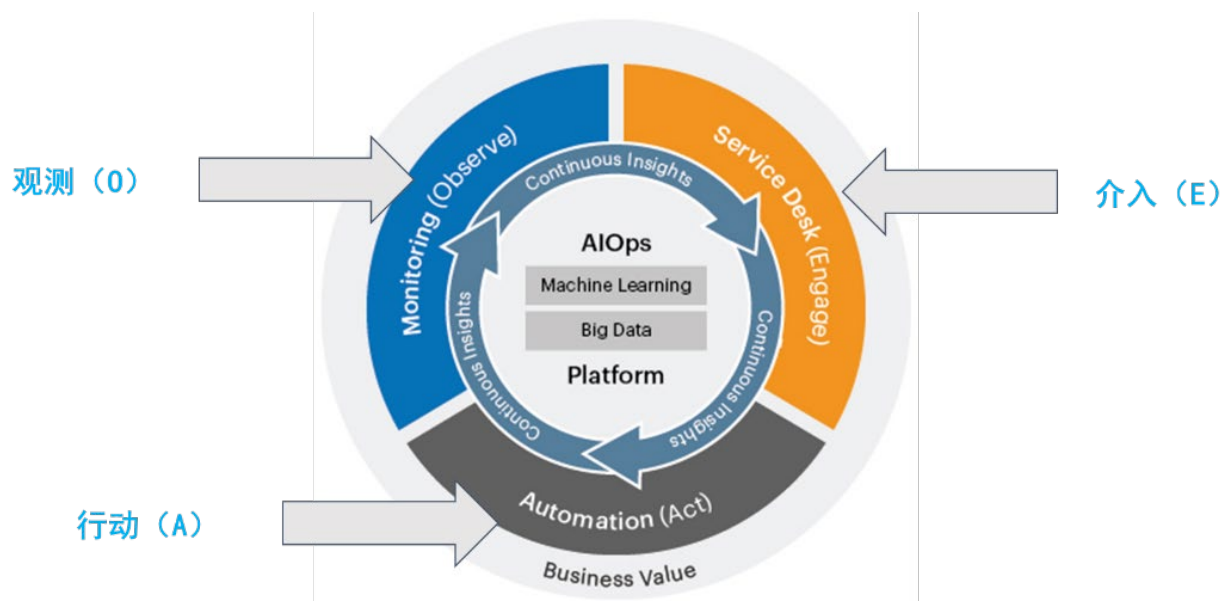
在实践中，基于机器学习的AIOps（智能运维），基本实现思路是要求组织利用大数据（BD）和机器学习（ML）功能来加速、自动化和简化 与保持系统性能高效、可靠运行 相关的所有任务。



注：ITOA单独利用大数据分析，而AIOps是使用ML和AI的实践演进。



“观测、介入、行动” (OEA) 循环



国际上对于基于机器学习的AIOps（智能运维）提出了，“**观测、介入、行动**”（OEA）循环的框架模型，OEA是一个概念，它代表了人工智能系统与其环境交互的循环过程，通常用于人工智能驱动的自动化和决策系统的上下文中。

以下是关于 OEA 循环中每个阶段的概述：

1. 观测

在观测阶段，人工智能系统从其环境中收集数据和信息。这可能涉及监控各种数据源、传感器、日志、用户交互或其他相关输入。系统观测和分析可用数据，以获得洞察并理解环境的当前状态。

2. 介入

一旦 AI 系统观测和分析了数据，它就会进入介入阶段，在此阶段，系统根据其分析和决策能力与环境或相关利益相关者进行交互，它可以根据预定义的规则、策略或机器学习模型提供建议、告警、通知或启动运维。

3. 行动

行动阶段涉及根据前一阶段产生的洞察和建议执行行动或做出决策。人工智能系统执行特定任务、执行运维或触发流程以实现预期结果。这些运维的范围可以从简单的自动化任务到涉及多个步骤

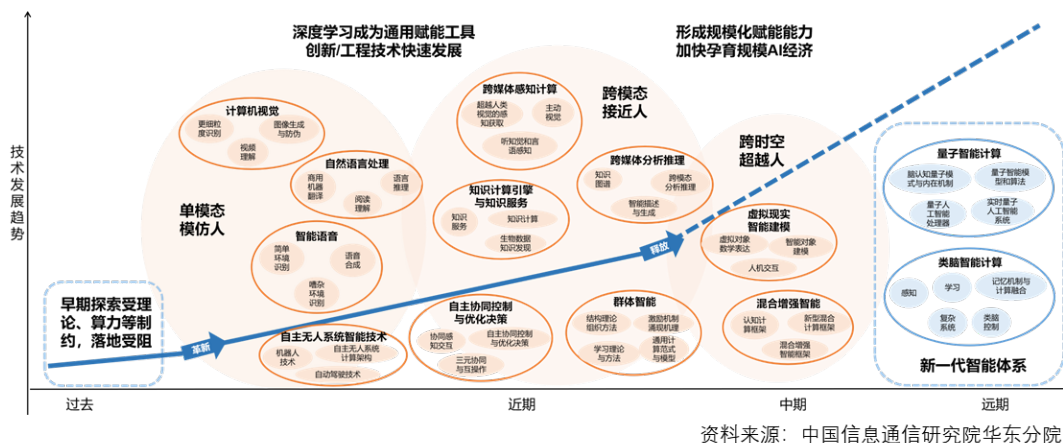
和交互的复杂工作流。行动后，循环从观测阶段再次开始，在各个阶段中不断循环以适应不断变化的环境并完善系统的响应。

OEA循环使AI系统能够随着时间的推移动态学习和改进其决策和自动化功能。

OEA环路通常用于需要实时或近实时交互的各个领域，例如自治系统、IoT（物联网）应用程序、客户服务聊天机器人或 AIOps 平台。它使人工智能系统能够有效地收集信息、分析信息并采取适当的行动，从而在动态环境中实现智能自动化和决策。

（资料来源：<https://research.aimultiple.com/aiops/>）

人工智能加速迈向AGI (Artificial General Intelligence) 时代



生成式AI: 基于现有数据生成全新内容的人工智能技术

大模型: 基于大规模参数的人工智能预训练模型

LLM: 基于深度学习的自然语言处理技术

ChatGPT: 基于LLM (语言大模型) 与RLHF (人类反馈) 技术的革命性产品

通用人工智能

- 生成式AI
- 具身智能
- 类脑智能
- ...

(资料来源: 中国信息通信研究院华东分院整理)

人工智能演化的三个阶段

阶段1: 狭义人工智能

- » 高度依赖监督学习算法, 需要大量人工标记的数据
- » 任务范围狭窄, 需要为新任务训练新模式
- » 确切知道人工智能可以或不能做什么

阶段2: 广义人工智能

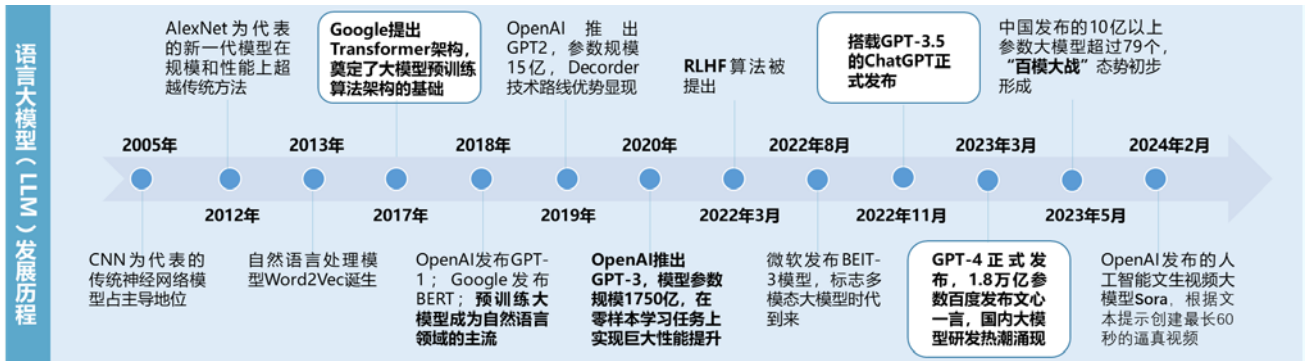
- » 自监督: 无需过多明示教学 (数据的去标签化)
- » 端对端: 一个模型可以完成多项任务
- » 判别工具 -> 生成式助手
- » 大模型: 涌现能力/零样本学习

阶段3: AGI (通用人工智能)

- » 比人类更聪明, 并且总是学着变得更聪明
- » 独立、自主发展...
- » 需要更好的治理、监管



大模型技术的特点和发展趋势



（资料来源：中国信息通信研究院华东分院整理）

大模型（大规模预训练模型）是指通过在大规模数据上进行预训练后能快速适应一系列下游任务的深度学习模型。主要技术特点如下：

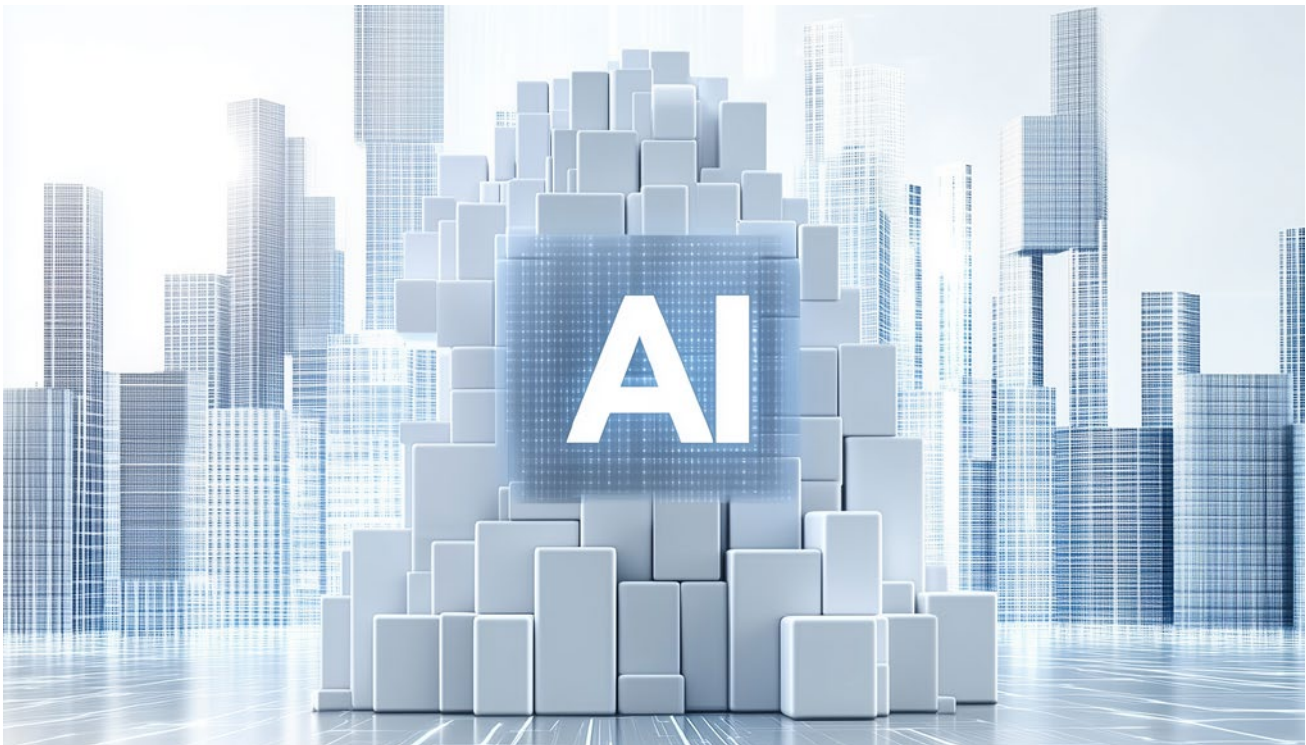
» **具有涌现能力：**在特定任务上，随着模

型规模提升，模型突然出现性能提升。

» **参数规模庞大：**参数规模不少于10亿（1B），严格意义上需要超过100亿（10B）。

» **具有通用性：**通过提示、微调，能够适应广泛的下游任务。

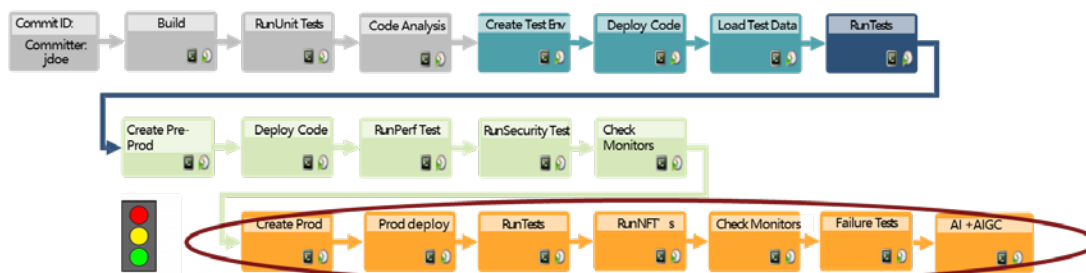
基于大模型技术的发展SRE专委会提出了更新版的OE模型：全面可观测+大模型+OE，以利于业内充分利用大模型等新技术赋能智能运维赛道。



SRE结合AIGC实现智能运维

SRE领导的智能化(AI能力的 “左移”)

SRE领导的服务自动化、智能化



在大模型为代表的新一代AI环境下，SRE聚焦生产环境（PROD）的AI应用—SRE结合AI技术，在各个运维场景实验和验证机器学习、大模型技术，深化和创新“生产智慧”，落实运维管理智能化，持续优化服务“稳定性”。

SRE社区专家关于人工智能运维的探讨

如何看待智能运维？

智能运维可以总结为，通过应用算法、模型和深度学习，针对特定场景进行智能运维。在这个过程中，数据的关键特征决定了机器学习的效果上限。然而，模型和算法只是用来逼近这一上限的技术手段。因此，在进行许多工作时，数据治理也是一个重要的工作领域。

我们可以将今年视为AIGC元年，目前AI业界已经成功实现了我们期望已久的人工智能工作模式，一些AI工具已能够进行一定程度的逻辑推理。

当前AIGC仍然存在一些限制，主流的概率性AIGC生成模型通常被认为是不可控的。在许多场景下，它面临着不确定性和幻觉问题，尤其是我们试图要求AIGC按照特定规则执行任务时，这种问题尤为显著。

AIGC到底带来了什么？

AIGC到底带来了什么？最大改变在于该模型具备通识能力和推理能力。过去，我们的故障分析主要依赖于数据的强关联，并利用这些数据关联的特征进行故障预测。而AIGC不仅能够收集这些隐含的关联关系，还能够结合白盒信息，例如将代码或业务特点传递给大模型，以便发现更广泛范围内的隐藏问题。这有助于我们快速建立智能运维体系。

我们在一些细分领域已经可以看到AIGC的能力对生产力的提升，此外有许多可探索的未知领域，而这些领域以前是不存在的。前者已经具备一定的确定性，例如代码生成、图像生成、机器人客服等；而后者则以多AI分工协作、个性化伴侣/助手等其他形态展开。

与此同时，AIGC在业务实践中也带来了许多突破和场景拓展。这些方面包括但不限于IT运维中的故障根因分析、运维知识库生成、故障处理方案以及方案的自动执行。此外，它可以帮助运维研发编写代码和实施工程，应用于测试领域中的测试方案和测试用例，以及更业务场景化的领域。

对于AIOps方面，目前语言大模型是基于概率模型的生成模型。这意味着它们在生成方案时具有潜力，例如将已发生的几种故障扩展成更多种类，并提示可能的指标和异常情况。然而，在使用AIGC生成文档内容或方案时，需要注意生成内容本质上是一个概率分布，其中可能包含许多有用的信息，但也可能包含冗余甚至错误的内容。因此，在使用时需要注意筛选并结合人工判断，以确保生成的内容准确和有用。



AIGC对运维或运维研发的影响是什么？

AIGC对运维或运维研发的影响相应的分为两类，偏落地改造既有流程的，可以通过CodeLlama/StarCoder等模型来实现代码辅助生成、通过微调模型来实现自动化的测试用例编写/漏洞检测/提交信息扩展等；而另外一类则有更大的野心，例如ChatDev/Autogen对于软件工程多角色Agent化的方式实现对既有软件工程的全部AI化，这部分工作对于复杂场景的效果目前还不太理想，不过很值得期待。

对于相对简单且逻辑上不太复杂的问题，例如磁盘告警或多次磁盘告警直接关联到写入失败的情况，AI大模型可以提供显著的改进。因此，在实际应用时，需要清楚语言大模型的能力边界，这将有助于更客观地评估其在不同场景中的应用潜力。

我们认为这在广义的代码生成领域也是一种形式，而其他领域，如测试代码生成，特别是在游戏开发企业中，提交代码时的提交信息可能不够详细，需要进行扩充，甚至需要对其提交的代码进行摘要，以确认其是否包含无效代码，或者其

提交的代码内容是否与提交信息一致。

在整个运维研发领域，AIGC应该非常擅长。至于运营方面，像企业文档和企业知识库，这正是当前自然语言大型模型的擅长领域之一，工作内容已经相当成熟，这里不再赘述这一点。

注：内容参考来源：

1、What is AIOps, Top 3 Use Cases & Best Tools? in 2024 ; Written by Cem Dilmegani

<https://research.aimultiple.com/aiops/>

2、中国信息通信研究院华东分院整理资料

LLM和Multi-Agent 在运维领域的实验探索

摘要

本文从智能运维面临的挑战和痛点出发，介绍企业运维领域应用 AIGC 的实践案例，基于确定性运维的实践经验，提出以 LLM 为中心，基于多 Agent 协同的运维方案，并提出在大模型时代下，对下一代智能运维的思考。



张曦

华为云AI使能专家，犹他州立大学统计学博士，研究方向为AI for Data、AI for BI、AIOps，时间序列分析等；具有丰富的人工智能在企业场景应用落地的成功经验，应用场景覆盖营、销、服、供、采、制、研发等多领域，曾支撑华为集团多个业务应用+AI，曾带领团队成功攻克5+企业技术难题，并主导发布多个AI服务。

自ChatGPT问世以来，AI迎来了奇点iPhone时刻，这一年来大模型深入影响企业办公、金融、广告及营销等很多领域，也给运维领域的挑战带来新的解题思路。我们洞察发现大模型给AIOps带来新机遇：已有云厂商利用大模型对运维事故进行根因定位并给出故障缓解措施建议，近七成以上运维人员对LLM的分析结果满意度大于三分。我们认为AIOps面临以下三大挑战：

挑战1：针对运维领域海量知识快速获取、辅助诊断和故障分析能力

在大模型如火如荼的时代背景下，运维领域应用LLM来获取运维知识，针对故障进行分析和推荐修复方案已是势在必行，如何将LLM较为广泛的知识储备（横向能力）与运维领域的专业知识（运维垂域）相结合，对具体故障给出较为准确和可以用作参考和修复操作的指令，是本方案的第一重要能力实现。



挑战2: 针对多模态数据进行快速高准确的异常检测能力

运维领域有丰富的多损益来源、多类型数据: 指标数据 (metric)、日志数据 (log) 以及调用链数据 (trace) 等多模态运维数据。指标能够反映业务状态和机器性能的时间序列数据; 日志是一种程序打印或执行代码输出的非结构化文本; 调用链则是在系统完成一次业务调用的过程中, 把服务之间的调用信息连接成的一个树状链条。多模态运维数据可以反映系统状态的全方位信息, 针对

不同模态的数据, 智能匹配最优算法得到较为准确的故障检测结果是本方案的第二重要能力实现。

挑战3: 针对多源复杂部署的运维数据进行快速根因定位能力

多模态运维数据一方面能够反映系统状态的全方位信息, 另一方面也给故障定位带来很大的困难。针对log、metric、trace三方面来源的告警信息进行有效、快速、准确地分析, 结合trace数据生成的实时拓扑结

构图, 排除故障传播节点, 给出故障根因节点推荐, 是本方案的重要能力实现之一。

在真实的运维工作中, IT系统何时发生故障、发生什么类型的故障都是未知, 只能根据以往运维经验和故障现象进行推断, 所以针对这一困难, 本方案利用LLM储备的相关运维知识进行few-shot tuning后能够给出初步的故障分类建议, 起到辅助运维人员进行故障分析和根因定位的作用。

结合以上问题我们进行深度探索, 并总结出一些方案, 在此, 分享给各位读者。

方案详述

本方案提出以LLM为中心结合多模态Agent协同的自主决策、自动修复的运维方案, 整体架构如图1所示, 我们重点强调三大关键技术:

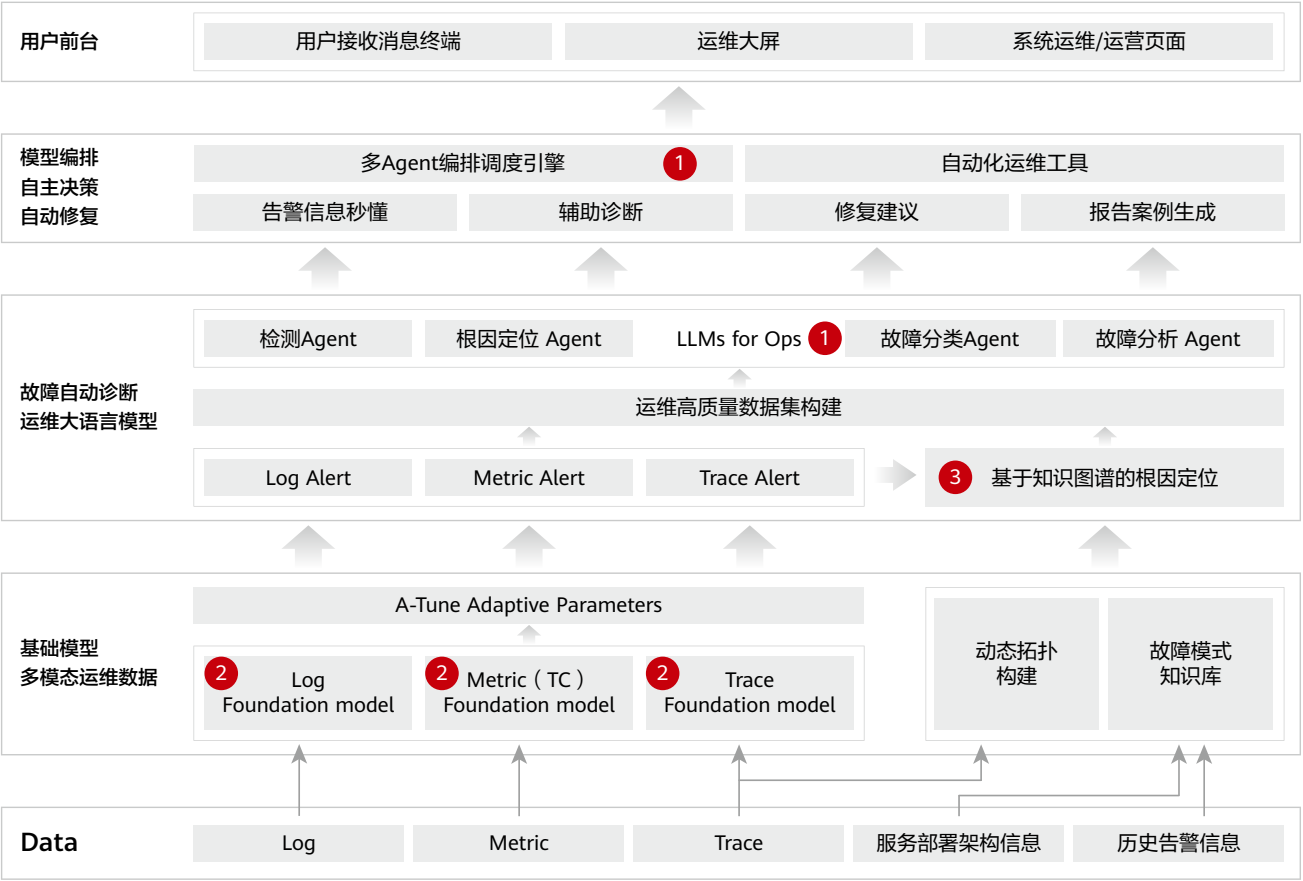


图1: 基于LLM和多模态Agent协同的自主决策、自动修复运维方案

关键技术1: 具备运维领域知识经验的LLM

- » 梳理异常检测告警信息和运维故障模式，形成运维领域高质量数据集，使用finetuning+外挂企业运维知识库的方案，使得LLM具备运维领域故障分析定位能力。
- » 使用主管LLM对异常问题与子领域Agent进行桥接，多个子领域Agent协同工作，提升运维效率。

关键技术2: 更全能的多模态数据异常检测基础模型

- » 针对成百上千条不同采用频率的metric序列采取一定维度聚合的方法，化繁为简，一旦检测出异常，再细化分析关键指标的关键表现。

- » 针对半结构化和非结构化日志采取针对性的解析方式：前者注重模板提取，后使用uADR/sADR进行异常检测，后者注重语义理解，使用BigLog预训练模型结合Deep SVDD+SAD进行异常检测。
- » 针对结构化明显的trace数据，抽取分离调用关系并定义节点关键数据，通过转化时间序列有效识别异常节点，并通过拓扑关系分析帮助推测可能根因节点及故障传播方向。

关键技术3: 基于知识图谱的根因定位

- » 利用DBSCAN从时间维度聚类，形成异常事件。
- » 利用Trace数据实时生成拓扑结构图。
- » 根据故障模式知识库分析故障传播链路。

针对Log, Metric, Trace三类运维数据我们分别构建了异常检测基础模型，一旦系统接受到异常即会通知LLM主管Agent，主管Agent对异常问题进行决策与子领域Agent进行桥接，多个子领域Agent协同工作，实现运维故障自动诊断和多个任务模型的编排，提升运维效率。

我们针对不同的数据源分别打造了三个基础模型，首先是metric异常检测，我们对不同对象不同采集频率的运维数据进行分组，然后根据指标变化曲线提取不同窗口的差分特征，第三步是将同类型的指标时序进行融合后利用多个异常检测器进行异常检测和时间聚类，一旦发现异常则通知LLM主管Agent进行决策，检测Agent收到详细检测的指令则进一步利用多个异常检测器进行细致的异常检测。

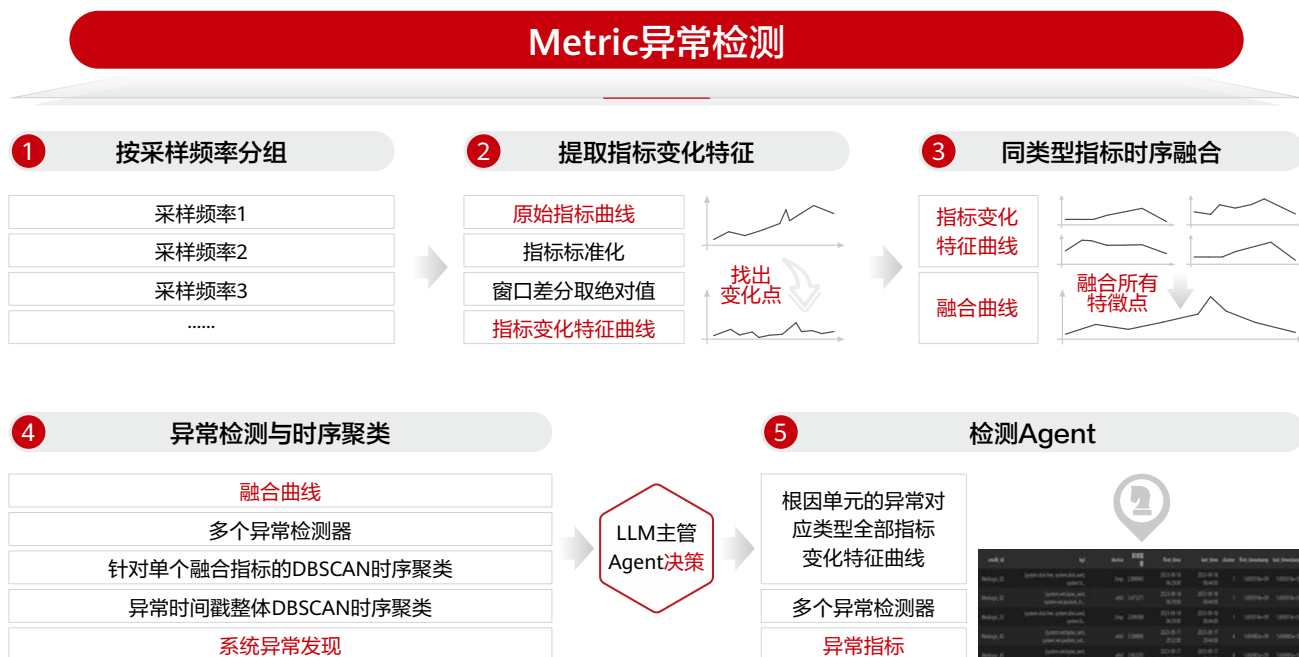
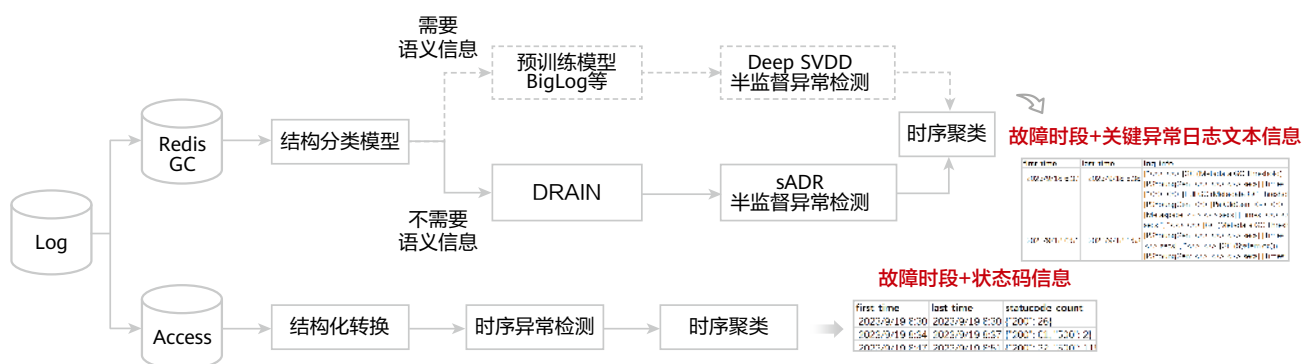


图2 多模态异常检测-Metric基础模型



- » **日志类型全覆盖:** 自动识别日志结构, 对结构化、半结构化和非结构化日志自适应匹配时序异常检测、基于日志解析提取模板的半监督异常检测和基于预训练语言模型BigLog理解语义的半监督检测算法
- » **高效率在线检测:** 可对流式日志数据实时检测并实时更新模型

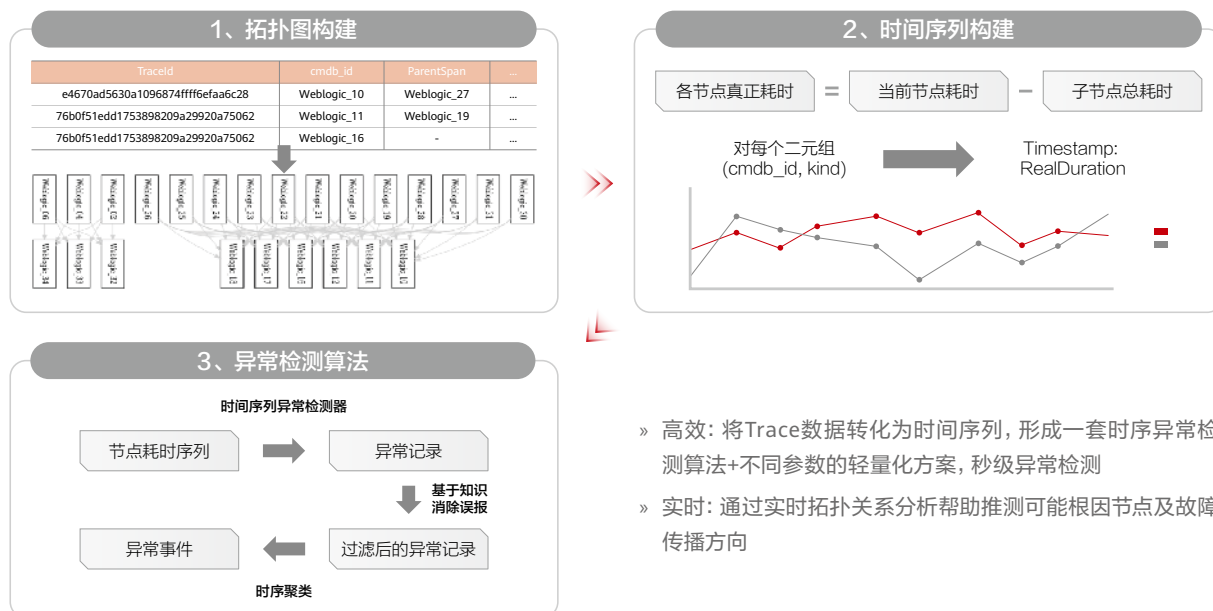
图3 多模态异常检测-Log基础模型

其次是Log异常检测, Log类型可以分为两大类: Redis GC和Access 日志, 针对Redis GC日志首先使用结构分类模型检测是否需要语义信息, 如果需要语义信息, 那么则会使用运维领域的预训练语言模型, 如BigLog等; 之后是由Deep SVDD进行半监督异常检测, 检测出不

需要语义信息的部分则使用DRAIN进行模板提取后, 使用sADR进行半监督异常检测; 最终对时间进行聚类后输出故障发生的时段和关键异常日志文本信息, 针对Access 日志我们使用结构化提取成时间序列后进行异常检测, 对时间聚类后最终输出故障发生的时段和状态码

信息。

针对Trace数据有两部分输出, 一部分是根据Trace信息实时生成动态拓扑作为根因定位的输入, 另一部分针对调用链节点之间的时长构成调用链时间序列进行异常检测。



- » 高效: 将Trace数据转化为时间序列, 形成一套时序异常检测算法+不同参数的轻量化方案, 秒级异常检测
- » 实时: 通过实时拓扑关系分析帮助推测可能根因节点及故障传播方向

图4 多模态异常检测-Trace基础模型

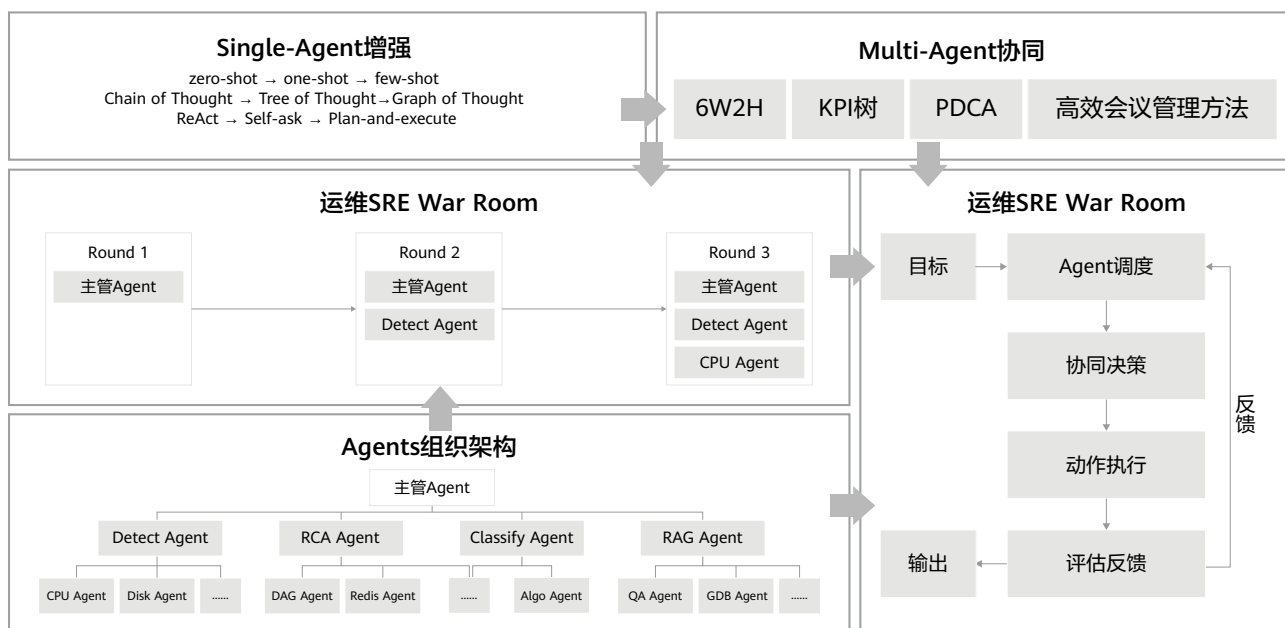


图5 多Agent协同框架：基于企业组织管理方法的高效协同诊断

结合大模型的技术演进趋势，从zero-shot->one-shot->few-shot，从Chain of thought->Tree of thought->Graph of thought，从Single-Agent增强到Multi-Agent协同，Multi-Agent在激发大模型能力的同时也带来了新的技术挑战，近期业界实践已发现，Multi-Agent架构如

果缺少合理的组织管理与协同沟通方法时，Multi-Agent的效果可能会比Single-Agent更差，我们认为在通过大模型通往AGI的道路上，需要将Agent当作平等的人类非工具来看待，因此多Agent的组织管理与协同需要借鉴管理“人”的方法，我们希望在基于Agents组织架构上，探索

一套基于企业组织管理方法的多Agent协同框架，例如通过KPI树方法分解任务，通过PDCA方法反馈循环提高Agent工作效率，通过企业高效会议管理方法来合理组织运维War Room的Agent进入和退出，实现多Agent的高效协同诊断。





根据详细检测的结果分析出根因节点为 Weblogic 16, 故障分类Agent根据根因节点信息及异常指标的描述异常程度判断为磁盘故障, 故障分析Agent这时会给出针对此故障的详细分析报告, 包含故障爆炸半径以及修复建议等。



方案创新性、通用性及实用性



创新

多Agent协同
完成复杂运维任务

- » 基于企业组织管理方法的多Agent协同框架，复杂运维任务处理更高效
- » 多Agent协同完成运维主流程：异常检测->根因定位->故障分类->故障分析->修复建议



通用

解决运维领域
共性问题

- » 多模态异常检测基础模型，包含Trace、Metric、Log数据全方面处理能力，开箱即用
- » 框架与算法不依赖具体特定应用场景，结合大模型实现较强的泛化能力



实用

故障恢复优先
各模块松耦合

- » 故障诊断报告体现可解释的故障爆炸半径，为实际生产运维故障快速恢复提供有力依据
- » 各模块松耦合可插拔，可以全面应用于各类场景故障快速恢复需求，已在公司内部多场景落地

最后，我们方案创新地提出基于企业组织管理方法的多Agent协同框架，让复杂运维任务处理更高效；使用多Agent协同完成运维主流程：异常检测->根因定位->故障分类->故障分析->修复建议；

构建的多模态异常检测基础模型，包含Trace、Metric、Log数据处理能力，开箱即用；框架与算法不依赖具体特定应用场景，结合大模型实现较强的泛化能力；故障诊断报告体现可解释的故障爆炸半

径，为实际生产运维故障快速恢复提供有力依据；各模块松耦合可插拔，可以全面应用于各类场景故障快速恢复需求，已在一些多场景落地实践。

参考文献

- [1] Zhang S, Pan Z, Liu H, et al. Efficient and Robust Trace Anomaly Detection for Large-Scale Microservice Systems. ISSRE, 2023.
- [2] Li D, Zhang S, Sun Y, et al. An Empirical Analysis of Anomaly Detection Methods for Multivariate Time Series. ISSRE, 2023.
- [3] Wang Z, Liu Z, Zhang Y, et al. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models. arXiv, 2023.
- [4] Jin P, Zhang S, Ma M, et al. Assess and Summarize: Improve Outage Understanding with Large Language Models. ESEC/FSE, 2023.
- [5] Chen Y, Xie H, Ma M, et al. Empowering Practical Root Cause Analysis by Large Language Models for Cloud Incidents. arXiv, 2023.
- [6] Zhou X, Li G, Sun Z, et al. D-Bot: Database Diagnosis System using Large Language Models. arXiv, 2023.
- [7] Zhou X, Li G, Liu Z. Llm as dba. arXiv, 2023.
- [8] Wen Q, Gao J, Song X, et al. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series. AAAI, 2019.
- [9] Liu Y, Tao S, Meng W, et al. LogPrompt: Prompt Engineering Towards Zero-Shot and Interpretable Log Analysis. arXiv, 2023.
- [10] Tao S, Liu Y, Meng W, et al. Biglog: Unsupervised large-scale pre-training for a unified log representation. IWQoS, 2023.
- [11] Ma L, Yang W, Xu B, et al. KnowLog: Knowledge Enhanced Pre-trained Language Model for Log Understanding. ICSE, 2023.
- [12] Zhong Z, Fan Q, Zhang J, et al. A Survey of Time Series Anomaly Detection Methods in the AIOps Domain. arXiv, 2023.
- [13] Wu H, Hu T, Liu Y, et al. Timesnet: Temporal 2d-variation modeling for general time series analysis. ICLR, 2023.
- [14] Yu G, Chen P, Li P, et al. Logreducer: Identify and reduce log hotspots in kernel on the fly. ICSE, 2023.

盘古大模型专业服务 使能千行万业

盘古大模型专业服务

行业AI咨询服务 | AI使能优化与提升 | 行业AI大赛与培训服务 | 华为云人工智能培训服务

TOP核心竞争力

数据质量高 | 模型行业丰富 | 工具易用 | 安全合规





乘云启航， 打造面向未来的数字化新零售商业体

—— i 百联业务中台全量上云实践分享

摘要

本文重点介绍百联全渠道业务中台全量上云实践经验，在面临集团业务多元性、架构复杂性及上云的复杂性的挑战下，通过上云试点、灰度切流、测试和演练等，确保业务系统顺利割接上云，乘云启航，打造面向未来的数字化新零售商业体。



王善良

百联全渠道电子商务有限公司副总经理，负责百联全渠道公司运维、数据安全、IT服务台等团队管理工作。

百联集团数字化转型历程

百联集团成立于2003年4月，由原一百集团、华联集团、友谊集团、物资集团合并组成。百联核心业务深耕上海、发展长三角，优势业务布局全国，涵盖主题百货、购物中心、大型卖场、标准超市、便利店、专业专卖等零售业态，经营汽车贸易、电子商务、仓储物流、消费服务、电子信息等领域。截至2023年底，全国零售网点近4700家，长三角零售网点占比超80%，上海零售网点占比约50%。其中，百货、购物中心、奥特莱斯门店数与总面积以及超商门店数均为上海第一，青浦奥特莱斯销售额连续多年蝉联全国第一。百联集团在2015年成立了百联全渠道电子商务有限公司，定位是紧密围绕商务电子化和传统业务改革转型两条主线，全力推进百联集团商务电子化。



2016年5月19日i百联平台正式上线,该平台上实现了百联集团内数张会员卡的会员打通、积分打通,完成整个集团下各业态会员一张卡的目标。从上线至今,i百联平台的功能定位从最早期的门店数字化改造及商务电子化平台建设,已经逐步完善深化到集团数字化转型的一系列项目中,并于今年8月16日正式完成i百联业务中台全量上云切换工作。

同时,在百联集团‘1+2+N’数字化转型战略中,上云是其中很重要的一个项目,以“上云”作为起手式,先业务上云,再云上创新,乘云启航,逐步实现企业数智化转型战略。

业务上云的挑战分析

百联集团业务架构多元复杂,i百联业务中台既包括了集团共享中台的内容,如:共通的会员、支付、促销等功能,也包括面向各二级公司业务特色的业态中台的内容,如:面向百联股份的到店业务,面向联华、第一医药的到家业务,同时平台也与多个外部合作商有系统功能交互。

公司上云面临的挑战主要来自于三个方面原因:集团业务多元性、架构复杂性及上云的复杂性。

» **业务方面:** 涉及百货、商超、物流、医药等,如事故对经济和民生影响大,社会面影响大。

» **技术方面:** 系统微服务之间网状调用,分批切换难度大,一把切换停机长、风险高。整个i百联业务中台涉及操作系统4000+个、数据库100+套、中间件70多套、微服务数600+个、定时任务数1300+个以及内部域名2000+个,上云迁移、适配任务识别及改造工作量巨大。

» **管理方面:** 整个上云迁移工作涉及集团内9家二级公司、公司内10多个一级部门及近300家第三方合作商协调工作,涉及多业态和多供应商沟通协商影响最小化的统一停机时间窗口难度较高。整个上云迁移方案涉及上云系统的适配改造、自动化运维平台改造、降级方案制定、降级演练、上云迁移演练、及配合上云验证等过程。上云方案过程中开展正式会议120+场,上云切换当晚近200+人现场保障。

业务复杂		技术复杂	管理复杂
线上业务	线下业务	IT数量大	涉及的部门复杂
全渠道 会员中心 支付中心 积分中心	停车场20+ 线下门店4000+	操作系统近4000个 数据库100多套 中间件XX套	集团内9+个二级公司 公司内10+个一级部门 第三方合作商300多家
业态多		应用微服务多	人员及项目管理难度大
集团内9+个二级公司		微服务数600多个 JOB任务1300多个 内部域名2000多个 测试用例4000多个	项目任务1000+项 正式会议120+次 切换当晚200+人 演练次数5次
三方平台多		改造及上云复杂	
300家三方平台		Runbook任务200+项	
• 业务涉及商超、百货、贸易、物流、医药等,对经济和民生影响大,社会面影响大		• 服务间网状调用,分批难度大,一把切风险高 • 上云迁移、适配改造工作量大	• 对多业态和多供应商的业务影响,停机窗口时间协商困难 • 多部门多干系人,协同工作量大

“三步法”业务上云方案

基于识别的业务上云挑战，百联全渠道联合华为云上云迁移团队制定了上云试点、灰度切流、测试和演练的业务上云三步骤方案。

上云试点方案

基于百联全渠道系统的复杂性及上云事故的巨大影响性，经讨论通过小范围试点，

验证迁移流程和方案的可行性，识别问题和风险，确保后续大规模迁移顺利进行。

试点目的如下：

- » 对上云方案进行验证
- » 进行团队协同磨合
- » 提前识别迁移风险
- » 通过实践积累经验

最终通过多批次小范围试点、总结和优

化，为最终全量上云切换打下坚实基础。

灰度切流方案

在搭建云上环境后，让应用访问原机房数据库，然后切少量流量到华为云生产环境以验证从网络层到应用层的部署正确性，应用层通过逐步灰度上云，验证无问题后将应用层和数据层全量切至华为云。

上云试点的方法

- 上云方案验证 - 团队协同磨合 - 迁移风险识别 - 实践经验积累	- 相对独立，关联关系简单 - 非核心业务，业务影响小 - 具有试点的意义：能验证某个方案或能验证上云的价值		- 识别问题 - 分析影响 - 改进措施 - 总结经验
确认试点的目标 ①	选择试点应用 ②	执行迁移小循环流程 ③	试点复盘总结 ④

上云试点举例

	试点目标	详细描述
上云方案验证	切换方案	验证 Nginx 灰度切流（1%，10%，50%） 验证应用跨云访问中间件和数据库 验证 WAF 切流量到华为云方
	自动化部署	验证使用 CMDB 一键式部署应用方案
	测试方案	验证内部 WIFI 接入灰度环境进行验证的可行性
团队协同磨合	沟通机制	6 类沟通会机制
	运作规范	切换规范、通报规范、评审规范、决策机制
迁移风险识别	跨云时延	灰度切流时，应用跨云访问延迟影响业务体验
	IP 白名单	上云后 IP 变更会影响业态和第三方业务
实践经验积累	迁移工具	使用 DRS 进行 MongoDB 和 MySQL 迁移
	云服务集成	应用访问华为云 OBS 适配改造 CMDB 与华为云对接



测试和演练方案

整个上云过程步骤多，需要进行充分的功能测试、性能压测和切换演练，为尽可能提前模拟上云操作步骤，基于华为云最佳实践1:1生产环境建立了切换演练环境。

搭建演练环境的目的如下：

- » 通过搭建演练环境，验证搭建的步骤/脚本/自动化发布系统，搭建云上生产环境出错概率低。

- » 在此环境可以进行上云演练操作。
- » 上云前在演练环境压测，可提前发现上云后可能存在的性能风险。

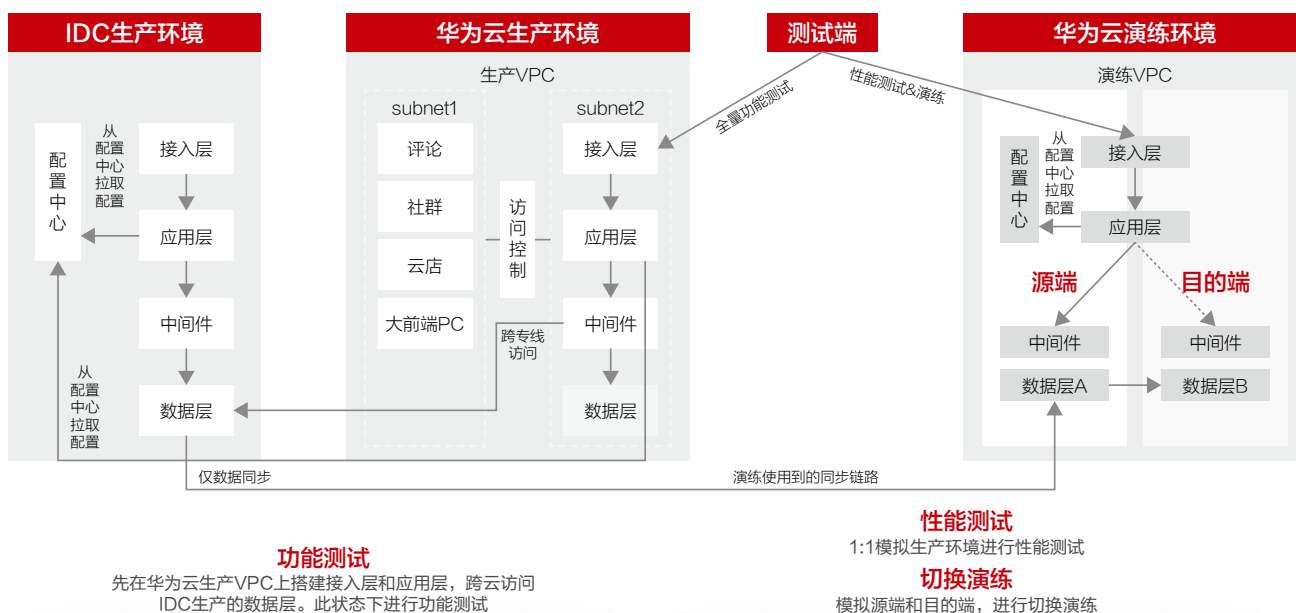
通过演练解决以下问题：

- » 通过演练执行多项切换上云步骤，提前熟悉操作流程，缩短切换停机时间。
- » 增强多人执行操作并行及串行操作间的配合默契。
- » 通过演练检查runbook遗漏点，演练

优化runbook，提高正确性和完整性，部分步骤通过脚本或工具自动化执行，缩短相关操作的时间，提升操作正确率。

- » 通过演练识别未知问题，减少切换当晚风险。
- » 通过演练加强技术积累，提升问题快速解决能力。

最终通过4正1回滚，做到切换上云问题收敛，将上云时间缩短6小时。



因为上云项目的复杂性，我们基于华为云上云迁移最佳实践制定详细的切换上云切换前Checklist和上云操作runbook。Checklist用于在切换前确保相关前置工作都已经顺利完成，runbook确保正式切换当晚现场200多人能够按照事先约定的步骤有序操作。

runbook每个操作精确到操作人、确认人、命令脚本、预计操作时长、串并行标识和回退决策点等。正式上云切换前CheckList共67项检查点，5次演练共执行300多个检查点，正式runbook共236项任务，5次演练共执行1000多项任务。

正是因为前期百联全渠道及华为云充分的准备工作，最终保障了百联全渠道比原计划提前完成上云切换工作，同时在上云切换成功后，能够一直保持整体运行稳定。



业务上云的经验及价值总结

百联全渠道庞大且复杂的业务系统最终顺利切换上云，得益于以下六个方面因素：

- » **清晰的战略目标**，引导上云过程中的决策和资源分配。华为云咨询服务协助百联制定了云转型的蓝图，并制定了三步走的战略规划，打基础、建能力、巩固与优化。
- » **成熟的方法论**，以华为云上云迁移方法论作为框架指导，确保上云过程的有序进行。
- » **项目早期的全局思维**，企业上云需要具备项目早期的全局思维，预则立，不预则废。在项目早期，项目组就思考整个上云的里程碑，需要在什么时间做什么工作，并梳理上云交付全景图。
- » **强有力的组织保障**，作为集团一把手工程，明确联合项目组成员分工，集团、业态、公司领导及各部门同事的大力支持、配合。
- » **高效的团队协作**，有效的沟通协调和决策能力是构建高效团队协作的关键。双方项目组认真严谨的态度，建立了坦诚的沟通机制、WBS跟进机制，鼓励思考和提出问题的解决方案的文化。
- » **完善的上云方案**，通过任务分解、方案先行、迭代试点、上云演练、复盘总结、经验沉淀的机制，保证上云方案的完美执行。



本次业务上云创造了三大价值:

- » **提高容灾便利性**, vpcpeer方便打通多机房。
- » **提升系统可用性**, 方便应用部署云多

机房。避免自然灾害、不可抗力因素导致的机房故障, 从而造成系统不可用。上云前DB异地灾备, 无应用高可用, 上云后应用整体双机房高可用。

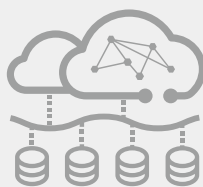
- » **加快业务创新**, 基于云上成熟PaaS产品和快速申请部署资源能力, 提升应用开发上线效率, 加快业务创新步伐。

提高容灾便利性



- » 无需租用机房和拉通专线, 同城AZ间高速免费内网
- » 无需容灾配置, 云服务原生支持同城容灾 (RDS、OBS、DCS、ELB等)
- » 容灾建设的**便利性提升**

提升系统可用性



- » 单IDC->多AZ, 由单机房可用性提升为同城多机房高可用, **可用性提升**
- » 计划外的**停机时间减少**

加快业务创新



- » 快速获得新技术进行业务创新, 如各种SAAS API、SDK等
- » 全球云基础设施, 支持贴近客户快速部署上线
- » 新特性的上市和**部署效率提升**

双轮驱动，构建运维“数循环”

一 数据驱动运维的方法与实践

前言

数据驱动运维是近几年业界的主流观点，但数据如何驱动运维？其方法及如何指导实践？尚未有清晰方案。本文构建了数据驱动运维方法论，提出“数据驱动运维”（DDOps: Data Driven Operation）的观点，构建运维领域的“数循环”，以数据为抓手打造完整的智能化运维实践方法论，为客户提供“道”、“法”、“术”一体的完整解决方案，希望对行业“用数据做运维”实践有所启发与帮助。



黄晓

云智慧首席咨询顾问，曾供职于国内头部上市软件公司，自2006年起从事IT服务管理领域，是国内ITIL思想最早的一批推广和实践专家之一，ITSS国标专家组成员，参与多项IT服务管理国标及行标的编制。工作经历覆盖IT服务领域产品研发、市场营销、售前推广、解决方案、业务咨询、项目实施、国标编制等领域。曾带领前公司团队发布多款IT服务领域自主知识产权产品，主导多个千万级IT运维项目交付实施。



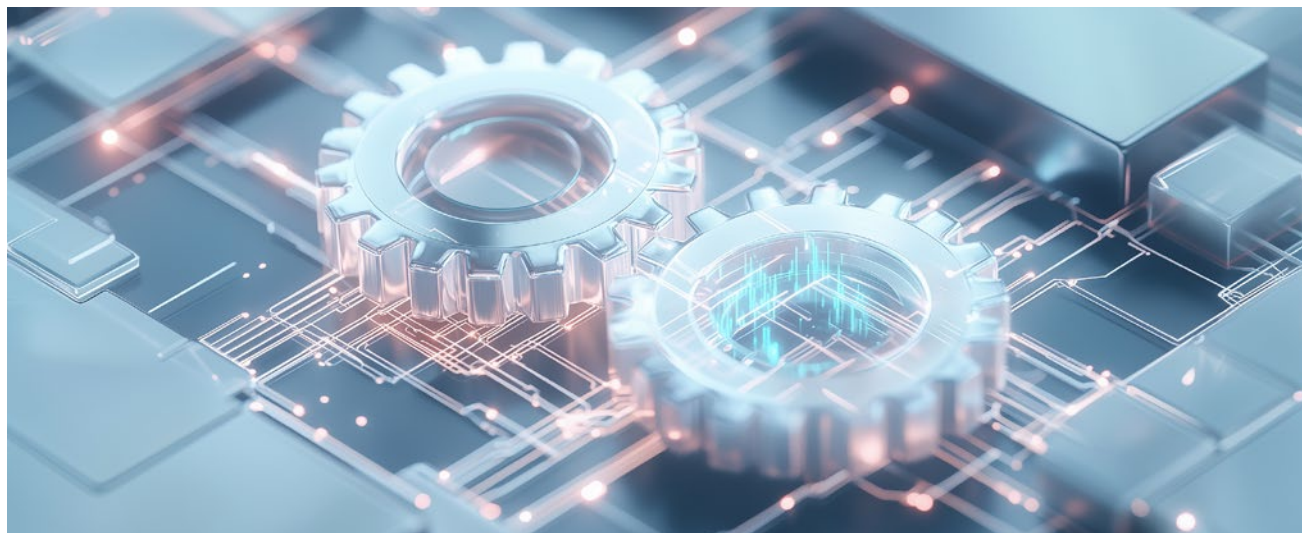
陆兴海

云智慧副总裁，目前负责咨询业务，具备十五年互联网、信息化以及运维相关领域产品规划、设计与研发经验，是国内IT相关服务领域最早的实践者和专家之一，智能运维国标核心编写组成员、工信部信创工委专家库成员。在作为云智慧产品负责人的七年内，主导与推动十几款运维相关核心产品的落地实现与营销、推广相关工作，帮助赢得了多个行业的核心客户认可，持有发明专利六项，著有《运维数据治理：构筑智能运维的基石》一书，合译《数字经济2.0：引爆大数据生态红利》一书。



伍杰

云智慧咨询总监，10年国内信息化咨询工作经验，在政府数字化转型、企业IT规划、IT服务管理、信息安全体系建设方面具有丰富经验。曾为数十家国内外用户提供咨询服务，涉及行业包括金融、通信、交通、政企、互联网等，同时也是国内智能运维符合性评估专家成员以及ITSS独立评估师。



核心理论：双轮驱动，构建运维“数循环”

对于数据驱动运维来说，其核心目标是构建运维领域的“数循环”，而不是孤立地看待如何管好和应用数据（图1）：



图1 运维数据“数循环”理念

更进一步探究数据驱动运维的核心内涵，IT运维数据是其逻辑底座，是实现“数据驱动”的基础，而构成数据驱动逻辑的内涵为：数据双轮驱动——数据驱动的两个轮分别是（图2）：

» 前轮驱动（聚焦目标），即价值牵引轮，负责引导IT运维的实践方向，通过指标数据度量运维价值，验证“养数”成

果，利用数字形成自顶而下的IT运维持续优化机制。

» 后轮驱动（聚焦建设），即建设推动轮，利用数据为IT运维的建设提供动力，驱动流程、平台、场景和运维模式蜕变。

数据驱动方法论总体框架着重于三大价

值目标，即提升综合治理、保障可用性、提高效能。其覆盖的运维模式包括事前预防、事中恢复、事后复盘以及其他新业态下的运维模式，确保可用性保障及管理运营两大运维场景落地。基于以终为始原则，通过双向数据驱动策略“运维数据治理”和“服务度量体系”来指导运维组织、运维流程、工具平台的建设及持续改进。



图2 数据驱动运维的核心逻辑：双轮驱动

“前轮驱动”：面向运维目标的持续优化（以用促通）

通过对IT服务价值度量的梳理，拉通业务、场景、流程与资源的端到端指标链，将指标数据作为抓手，用数字说话，形成自上而下的持续优化体系，打造数据驱动的IT运维持续优化能力（图3）。

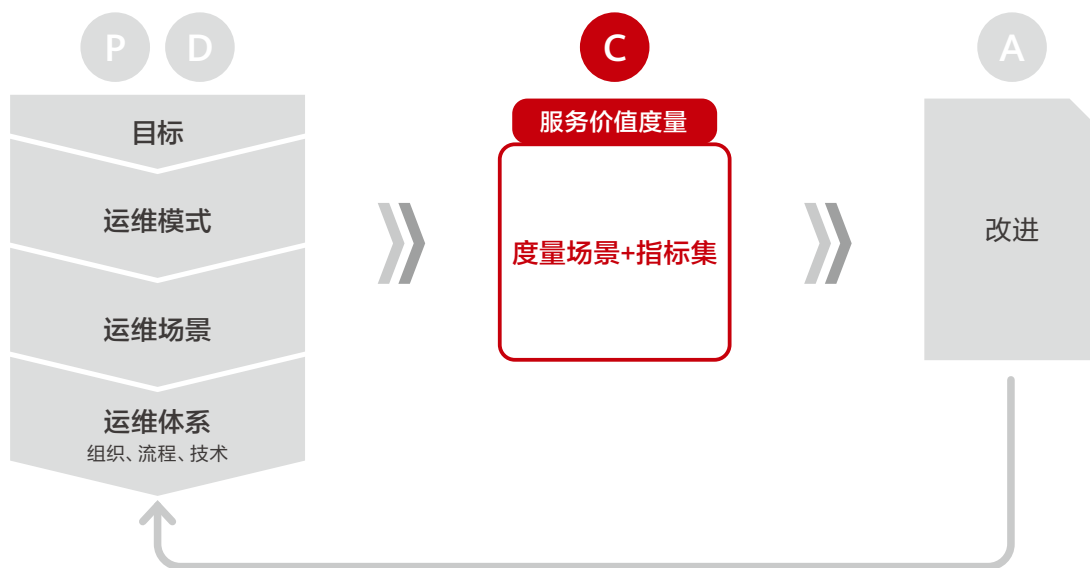


图3 前轮驱动：面向运维目标的持续优化



数据驱动运维“前轮驱动”的核心抓手是IT服务价值度量，IT服务价值度量是通过收集和分析相关数据来衡量和评估IT服务的质量、效率和效益的过程。IT服务价值度量的重要性在于它可以帮助组织了解和评估其IT服务的实际效果和价值。通过收集和分析相关数据，可以获得关于IT服务的各种指标和度量结果，如服务可用性、平均故障修复时间、客户满意度等。这些指标和度量结果可以帮助组织了解其IT服务的优劣，从而指导决策和改进。数据驱动运维强调基于数据的决策和实践。通过收集和分析相关数据，可以获得客观的、量化的信息，而不仅仅是主观的意见或直觉。例如，在面临故障修复决

策时，可以通过分析历史数据了解不同解决方案的效果和成本，并选择最优方案。IT服务价值度量需要依赖数据来进行分析和评估，而数据驱动运维则需要IT服务价值度量来提供支持和指导。通过结合IT服务价值度量和数据驱动运维，组织可以更加科学地管理和优化其IT服务，提高效率和质量，降低成本和风险。

结合国内外的相关研究，笔者提出了运维价值度量体系（简称OVMS，Operation Value Measurement System）建设的模型，主要包括总体框架、五大价值维度、全栈全生命周期覆盖、指标体系建设方法、价值度量的实施方法等。

运维价值度量体系建设的总体框架如下图4:



图 4 运维价值度量体系(OVMS)建设总体框架

该总体框架以支撑组织的业务战略和实现组织的业务价值为目标, 基于国内外的一些成熟理论和模型, 结合运维领域的实际需要进行逐层构建。关于运维价值度量体系可参考《华为云确定性运维专刊第二期》笔者所著《运维价值度量体系建设方法与实践》一文。

“后轮驱动”：驱动运维的五大转变(以通促用)

1 IT和业务深度融合, 运维模式转型

数据驱动IT运维模式从“被动 → 主动、离散 → 集中、人工 → 智能、模糊 → 可观测”转变。IT运维模式转型的四层内涵包括:

(1) 被动 → 主动

主动预防性运维, 改变传统的响应式、灭火式的运维模式, 转为主动预防的模式, 将避免重大故障、降低整体故障率作为核心目标开展运维工作, 如开展: 系统可用性持续强化、风险级应急管理、故障趋势预测等。

数据驱动逻辑: 主动预防性运维的工作开展需要建立在高质量数据的基础上, 以下以系统可用性强化活动举例。可用性持续强化包含以下几个步骤:

- » 系统脆弱项识别
- » 系统强化方案制定
- » 系统强化实施
- » 系统强化效果验证

其中第1步、第4步离不开数据的驱动

第1步 系统脆弱项识别: 需系统架构数据、部署数据、节点配置数据、近期性能及负载数据等。

第4步 系统强化效果验证: 基于监控数据的汇聚, 评价系统强化后的效果是否达成预期。

(2) 离散 → 集中

一体化运维, 以数据融合为基座, 构建场景导向的一体化的运维模式, 实现合作型团队向协作性团队的变革, 如: 一体化监控、研运一体、监管控一体等。

数据驱动逻辑: 其一, 数据实现了对运维场景、流程、工具平台的改造, 将影响IT运维组织的工作模式, 会促成IT运维组织



由传统以职能线条为主，线条间的合作模式往以企业业务目标导向，IT运维价值流为主线的跨职能协作模式。

其二，通过数据融合，将运维团队的关注点从各自职能范畴内的业务，上移至以价值场景导向的融合业务，各团队针对同一“运维实践”实现工作界面共享、运维数据共享，改变运维团队间的沟通机制。

(3) 人工 → 智能

智能化运维，以自动化和智能化为手段替代传统人工运维方式，获得更好的质

量、效率、成本，如：故障趋势分析、故障自愈、自动化发布、自动化部署等。

数据驱动逻辑：智能化运维的核心要素包括：组织、流程、资源、技术、数据、算法、知识，其中数据又是支撑其他6个要素的基础，所以智能化运维离不开数据驱动。

(4) 模糊 → 可观测

可观测运维，IT运维可观测不仅仅是通过监控的手段实现IT运维对象状态可观测、性能可观测，还包括IT运维其他管理领域，如：运维场景效果可观测、流程效能可观测、组织绩效可观测等。

数据驱动逻辑：可观测运维的核心逻辑是将观测对象的各维度情况，以数字化的方式呈现给对应的观测者，而这些数字的输出均源自各领域观测资源数据，以及这些数据的运算结果。

2 增强与创新IT运维价值场景

数据驱动运维，一方面增强了传统的运维单场景，一方面创新衍生出新的运维融合场景，对价值场景的驱动，能够更深度地解决复杂的运维问题。在数据运维方法论中，将运维场景划分为“可用性保障”和“管理运营类”，如下图5所示：

运维场景（融合）						
可用性保障场景			管理运营场景			
事前预防类	事中恢复类	事后总结类	服务管理	资源运营	客户运营	改进优化
系统持续可用性强化	故障诊断	故障证据链收集	请求响应及实现	资源容量管理	客户关系管理	组织效能优化
风险应急演练	故障影响分析	故障复盘报告生成	服务发布及推广	资源容量调拨	客户满意度管理	流程效能优化
变更风险控制和验证	解决方案制定	同类型故障横展排查	需求收集及反馈	资源分发回收	客户投诉、反馈	工具平台优化
故障趋势预测	故障处置	问题根因及方案知识化	服务级别管控	资源成本管控	...	供应商效能优化
日常值守巡检	故障恢复确认	预案迭代优化	...	...	...	...
...	应急指挥调度...	...				

图5 运维场景大类

为了让读者更好理解,下图以IT团队职能的视角划分维度,将这些场景又分为单点场景和融合场景两类(图6):



图6 IT运维价值场景地图(IT团队职能视角)

(1) 数据驱动单点场景: 通过跨领域数据融合, 为原生单点场景提供更多增益数据, 强化单点场景的业务深度。例如离散告警场景: 该场景告警由单一监控工具发出(暂时不提统一告警的模式), 各监控工具因为没有获取监控对象的相关归属信息(归属方、维护方、责任人等), 无法动态获取告警的通知对象及联系方式, 告警的通知规则的接收方只能设置为静态的电话/邮箱/IM号, 当人员组织架构发生改变后, 这些规则的维护工作很容易被忽视。而在数据融合的基础上, 即使不做统一告警服务, 单监控工具也能获取到监控对象的相关权属信息, 实现动态通知规则的配置。

(2) 数据驱动融合场景: 数据的融合为融合场景实践提供土壤。融合场景是指跨多工具平台、跨不同职能领域的运维场景, 例如: 全链路可观测、一体化监控、统一告警、重保综合支持等。这些场景需要跨领域、跨工具、跨团队数据流转和集成, 才能具备上层业务的融合。例如一体化监控场景: 一体化监控将不同分层管理对象的监控信息(网络、设备、中间件、数据库、交易等)、管理对象的近期变更信息、日志信息融合在一个监控平台里, 为跨职能团队的用户提供统一的工作界面。这种监控模式的实现的核心工作就是数据的标准化、融合及应用。

3 场景导向的IT运维流程再造

流程再造不是要颠覆已有的运维流程, 而是站在运维场景的视角, 以场景的价值目标导向, 对已有的流程进行重新审视, 再进行重构和融合。

基于数据驱动运维的方法论, 传统的“运维流程网络模式”可以向“场景化实践模式”转变。IT运维数据的融合为IT运维流程从传统的“运维流程网络模式”向基于价值场景的“场景化实践模式”的转变提供了基础和驱动力(ITIL 4的核心之一)。“运维流程网络模式”示例如下图7,

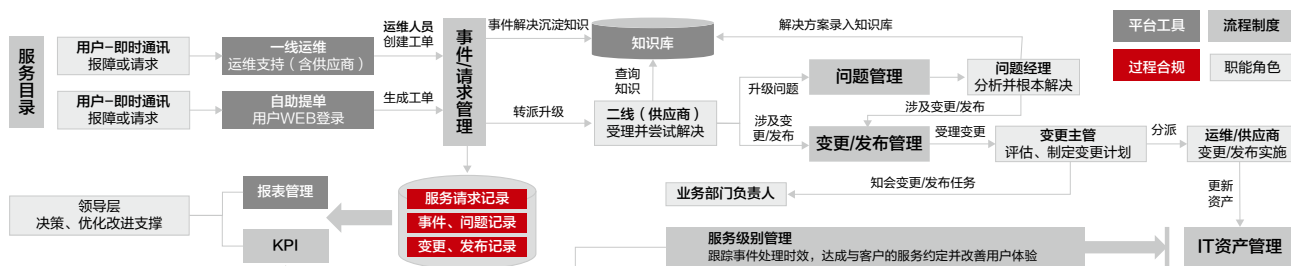


图7 传统的运维流程网络示例

该模式注重流程本身而不是全局的价值链，忽视了虑价值和场景。

“场景化实践模式”的典型特征是，某个实践域中由多个流程、多类人员、多个组织参与其中，将他们凝聚在一起的是价值目标和价值链，而价值链的落地本质上是数据流的实现。

这里以“可用性保障”的流程再造为例进行说明。流程再造并不是指删除并重新设计已有的运维流程，而是站在运维场景的视角，以场景的价值目标导向，对已有流程进行重新审视，再进行重构和融合。

以“可用性保障”类场景为例，针对可用性保障，可将整体的价值目标拆解为：

- » 事前 —— 降低故障率（避免重大故障发生、减少发生概率）；
- » 事中 —— 提高故障恢复的效率，提高故障恢复的有效性；
- » 事后 —— 减少同类故障的再次方式，沉淀经验和知识。

若以传统的“流程为中心”的思维，会粗

浅地将这些内容对应到“故障”处理流程，而以场景导向的思维，会将这个课题先转化为以下的场景蓝图8：

基于该蓝图，原有的流程，如故障、连续性管理、变更管理、应急管理、监控和告警等将面临以下的变革。

1) 流程可能是会跨多场景存在的。例如上述地图中故障流程将会跨事中和事后两类场景，在这种情况下，对于故障流程的设计不再局限在流程图和角色对应更多是在不同场景下，干系人关心什么、想看到什么、想联动哪些信息、有哪些任何组织要进行信息共享。

2) 单场景会包含多个流程。例如事前类的场景，会包含变更、连续性管理、SRE工作流程、建转运维流程等内容，如果按传统的流程为中心的思路，以流程先行的设计，很容易出现场景内流程设计缺失或不足的情况，导致场景目标的达成打折扣。

3) 流程间的关联不再是简单的数据互相引用。例如变更流程与其他流程的联动，

传统的流程为中心的模式，变更流程与其他流程联动的落地方式通常是简单的数据引用，会忽视为什么要联动、什么时候要联动、联动的时候有什么数据要进行传递、联动的过程什么样的。如变更后导致的信息系统的架构发生变化，那么这个系统原有的风险库、应急预案以及配套的计划性日常保障任务是否也应同步？如果要同步后续工作的Owner是谁？同步的结果有谁怎么确认？这些都是场景导向的流程设计会考虑，但流程为中心的设计会忽略的问题。

4 运维模式变革推动IT组织架构优化

数据驱动运维模式转型以及数据治理、智能运维带来内外部协作以及管理变革作用下，推动IT组织架构也需要优化。

其一，数据驱动本身涉及一些工作，如数据治理、指标体系需要有具体的人员进行对应。另外整体IT团队也会逐渐强化对数据的敏感度和认知，人员能力象限也应发生拓展和改变；

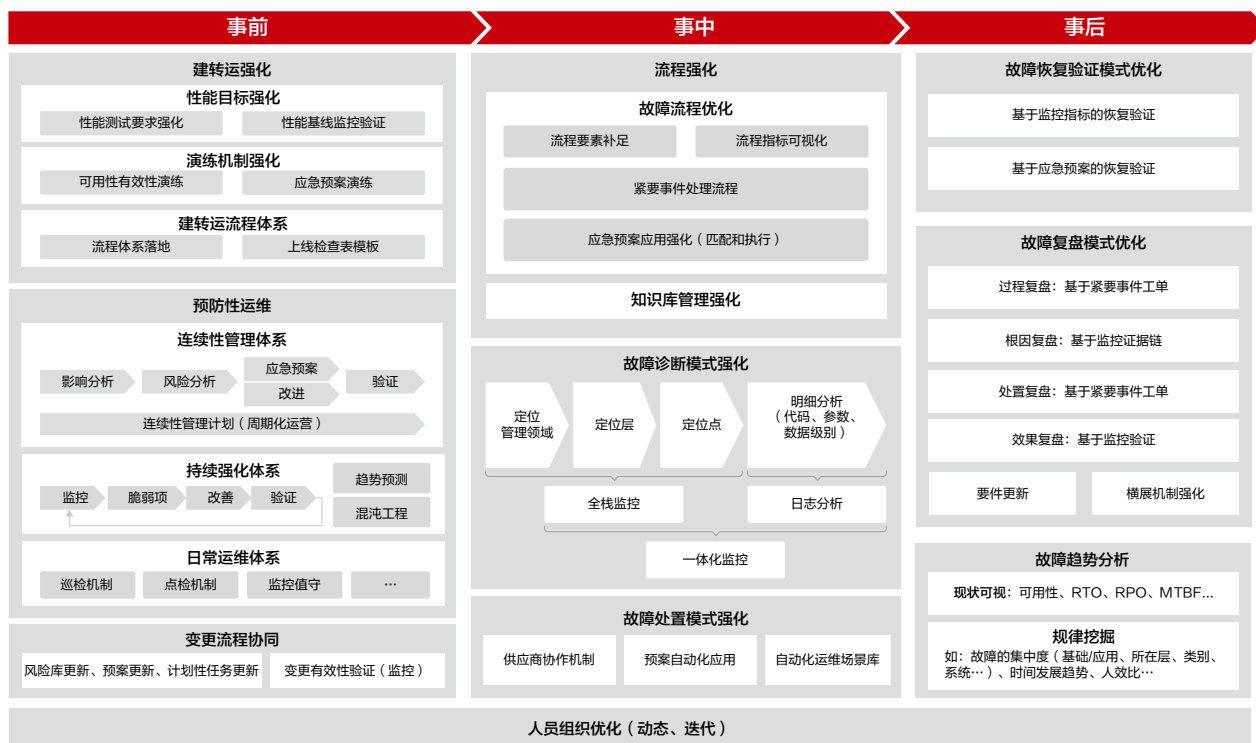


图8 可用性保障场景蓝图

其二，通过数据驱动的运维模式转型，重新定义各团队职责、边界以及协作模式，优化已有人员组织模式；

其三，通过数据驱动带来的运维技术变革，势必对运维效率、质量带来提升，原有的人员岗位职责以及人效分配逻辑也应随之改变，同时，如何改变以及改变后端效果评价都离不开数据驱动；

其四，通过数据驱动运维模式的持续迭代，为团队培养更多一专多能、多专多能的人才，进而推进IT运维组织架构变革。

5 数据统一融合助力IT运维平台重塑

IT运维数据治理的核心成果之一是“IT运维数据的集中和融合”，对IT运维平台能形成以下驱动：驱动统一的、集成式的IT运维数据的构建，为高端IT运维融合场景提供融合数据，反向驱动IT运维平台的进一步融合重塑（图9）。

在数据驱动的背景下，软件工具从“离散”的运维工具集走向融合型的一体化运维平台，是已经发生的事实。与以往的“离散”相比，这种“统一”表现在以下几个方面。构建统一运维管理平台，提供统一门户，实现统一用户管理及登录、工具模块的平台集成，通过统一数据采控实现已有和未来多种运维工具的集中式数据采集及其调控通过搭建统一的运维数据管理平台实现运维数据的集中存储、分析、指标体系的构建与管理以及数据质量和安全监管，通过统一告警事件管理的构建实现各类监控工具的告警事件消息归集、告警事件的抑制以及规范化管理，通过统一监控管理构建实现各类监控数据的高纬度分析、实现多维度数据高效联动，通过可视化管控中监控大屏与数据报表能力的构建，帮助各类用户基于数据进行快速分析与决策，通过构建统一的运维服务管理帮助组织建设高效规范的流程体系与服务模式，提供高质量、高效率的IT服务。

1) 统一运维门户：应提供统一的运维门

户，实现组织、用户与权限集成、多系统单点登录以及数据采集、数据处理、数据分析和展示。同时还应提供统一身份管理中心，实现统一用户管理、统一权限管理、统一认证管理和统一审计管理四大核心功能，实现组织机构和用户的全生命周期管理。

2) 统一数据采控：应具备运维数据统一采集管理能力，实现运维数据的集中化采集管理与调控，负责从运维工具或系统中通过各类协议或方法采集运维元数据、性能指标数据、事件数据、工单数据等。

3) 统一数据管理：应建立统一的运维数据管理，负责对采集的数据进行过滤、清洗、加工、存储、统计计算、数据建模、发布数据API等，能够对数据资产统一进行管理，满足对数据应用的敏捷开发、智能分析、数据血缘、数据地图、数据质量、数据服务等能力，能够充分利用数据，可实现复杂业务的数据开发任务，支持多种计算和存储引擎服务。

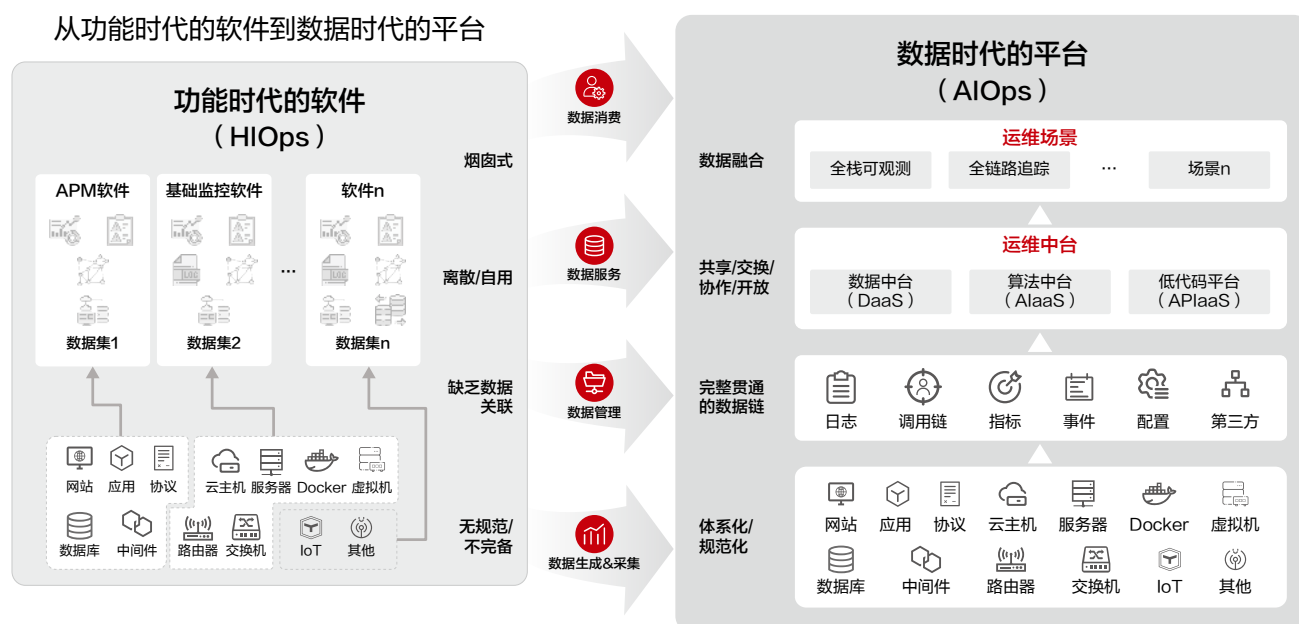


图9 离散运维走向集成式的统一运维

4) **统一事件管理**: 应实现统一的运维事件管理, 对来自各种监控系统的告警消息与数据指标进行统一接入与处理, 支持告警事件的过滤、通知、响应、处置、定级、跟踪以及多维分析, 利用多种算法实现告警事件的收敛、异常检测、原因分析等智能运维场景化应用, 实现问题事件全生命周期的全局管控。

5) **统一监控管理**: 应提供统一的运维监控管理, 结合大数据分析及人工智能等功能, 监控范围覆盖网络设备、服务器、

应用系统、虚拟化、存储、动力环境、安全设备等, 通过收集业务日志、监测指标以及业务链路关系等相关数据进行融合, 构建业务拓扑链路, 实现全资源、综合、立体、可视化的实时监控分析, 以关系串联指标、日志、告警、属性、变更等数据, 结合算法对业务、应用、服务、基础设施进行全资源统一监控、统一分析, 帮助快速发现、分析并定位故障。

6) **统一可视化管控**: 应构建统一的可视化运维管控中心, 提供各种数据可视化

方式包括可视化大屏与数据报表等, 通过可视化展示系统的运行状态、性能指标、异常情况等信息, 帮助运维人员更好监控和管理系统, 快速发现和解决问题, 提高系统可用性和稳定性。

7) **统一运维服务管理**: 应构建运维服务管理, 为组织提供高质量、高效率的IT服务, 整合企业或组织的所有IT资源, 包括硬件、软件、网络、安全等资源, 通过标准化、流程化的运营方式来提供IT服务。

“后轮驱动”核心抓手: 运维数据治理

运维数据治理是构筑数字化运维的基石, 需要继承和借鉴传统业务数据治理的理论方法, 针对运维数据的特点, 明确运维数据的治权和治理框架, 形成运维数据治理体系, 提供高质量、全覆盖的运维数据支撑数字化运维场景更好落地实施。

数据治理是一个复杂的工程性工作, 涉及大量资源投入, 需首先明确运维数据治理不是为了治理而治理, 其核心意义需要从运维价值驱动, 即: 控制IT风险、提升交付速度、提升客户体验、提高IT服

务质量。在此价值创造的基础上, 模型提出了运维数据治理的目标: 获得更加准确、好用的运维信息资产。这个目标有3个关键词: “准确”的数据是智能化运维的基础, 数据不准将导致智能运维场景不可用; 智能运维是一种全新的运维模式, “好用”的数据将有助于智能运维的应用, 数据应用又能反过来提升数据的准确性; 运维数据类型很多, 要用好数据, 需要数据达到“信息资产”级别。

要让运维数据转变为信息资产, 需要围

绕治理方法、治理过程、技术平台3要素, 持续完善运维数据治理。在治理方法上, 要关注以运维指标体系为代表的主数据管理, 以CMDB为代表的广义元数据管理, 并基于数据标准、质量管理、安全管理形成运维数据治理的关键治理工作。在治理过程上, 需要借鉴PDCA、IT治理、精益创新等思路, 重点划分为策略、建设、运营三个闭环。在治理工具上, 应建立存量与新建工具组合的工具支撑, 包括运维数据平台、指标体系、CMDB、监控、数据门户等工具。





大模型与智能运维的双向奔赴

背景

本文主要介绍在大模型智能运维场景下，通过大小模型的结合，结合系统运行状态进行概括、溯源、推理根因，输出故障状态摘要和根因诊断结果，最终实现故障检测能力和运维效率提升。



陈鹏飞

中山大学教授、博士生导师

在当今科技发展的汹涌浪潮中，云原生技术与大模型的融合正以前所未有的速度重塑着各个领域的格局。随着数据量的爆炸式增长、计算需求的日益复杂、人工智能技术的深入发展及业务场景的不断拓展，基于云原生大模型应运而生，成为推动技术创新和产业升级的关键力量。在这一背景下，智能运维能力显得尤为重要。

云原生已成为支撑大模型的基础技术，使得模型的训练和推理更加高效。同时，智能运维实践可以帮助企业更好地理解和管理这些复杂的系统。

通过分析自动发现IT系统中的问题，并定位根因，自动恢复系统，增强IT系统稳定性、性能和可用性，为IT运维提供“端到端”的解决方法。

大模型能够基于深度学习技术，自动识别运维中的异常模式，而智能运维则能依据大模型的分析结果，迅速制定并执行相应的应对策略，实现快速响应和高效解决问题。两者相互赋能，形成了一个良性的循环。大模型为智能运维注入了强大的智慧力量，智能运维则为大模型提供了丰富的实践经验和数据支撑，共同推动运维领域向更加自动化、智能化的方向发展。



大模型赋能智能运维

大模型成为智能运维的指挥大脑,与其他智能体协同工作

在这个过程中，终端用户、SRE、运维机器人和数据中心发挥各自的价值与协同作用，这种协同工作不仅提高了运维效率，还能够实现更高级别的自动化，为企业带来巨大的经济效益。

» **终端用户将通过与运维机器人的交互，实现对系统的实时监控和故障排查。**大模型可以通过分析用户行为和反馈业务情况，预测和诊断潜在的问题，如用户反映“业务变慢：为什么请求变慢？”等问题，直接反馈业务需求

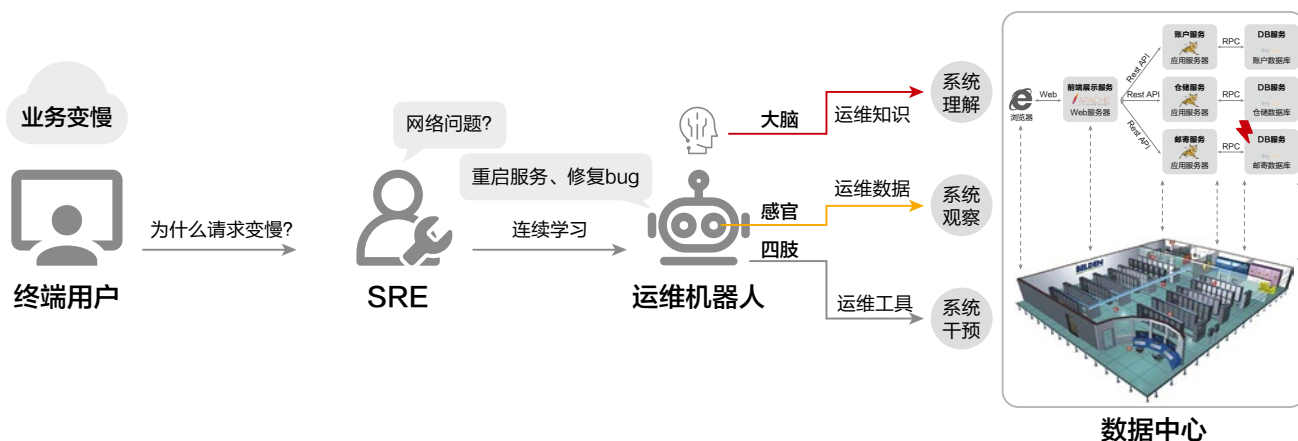
到SRE角色，SRE结合大模型能力，进行分析和判断，进行下一步的动作。

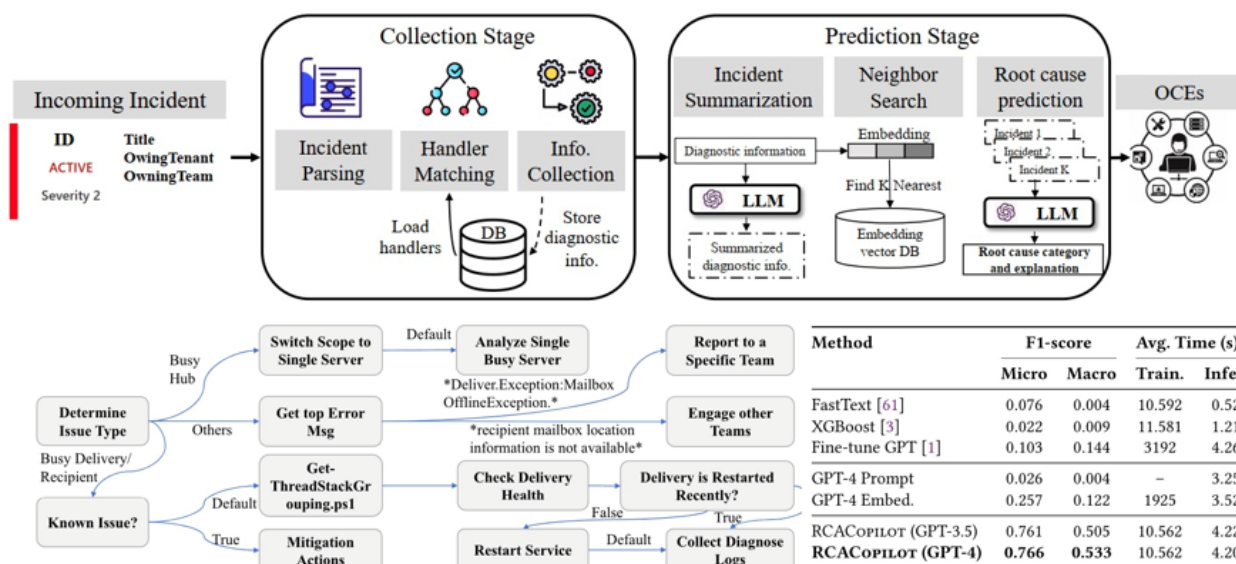
SRE将作为大模型的重要组成部分，通过对系统的深入理解和优化，实现对系统的持续改进。大模型可以通过对系统性能和可用性的实时监控和分析，发现和解决潜在的故障和瓶颈，提高系统的可靠性和稳定性，从而保证业务的连续性和可持续发展。

运维机器人将通过大模型的分析 and 决策, 实现对系统的自动化运维。运维机器人将大模型作为“大脑”和可观测工具作为“感官”, 能够感知环境并做出

相应的决策。当系统出现故障时，机器人可以通过“感官”获取相关信息，结合大模型进行分析和决策，通过“大脑”对运维知识进行系统的理解，并对运维数据进行系统观察，同时借助运维工具进行系统干预，最终实现对系统的自动化运维。

大模型也将通过对数据中心的分析和预测，实现对系统的优化和改进。通过对数据中心的大量数据进行深度学习和分析，大模型可以预测未来的需求和趋势，对数据中心的分析和预测，可以帮助数据中心实现对系统的优化和改进，提高数据中心的运行效率和稳定性。





基于大模型的云故障根因定位分析

通过大模型做到智能运维如何通过大型语言模型对云事件进行自动根本原因分析。

- » **数据采集与整合:** 将云环境中的服务器日志、监控指标、配置文件、应用程序日志、用户操作记录等各类数据, 进行整合和预处理, 使其能够被大语言模型有效读取和理解。
- » **模型选择与训练:** 选择合适的序列数据和文本数据的大语言模型架构, 使

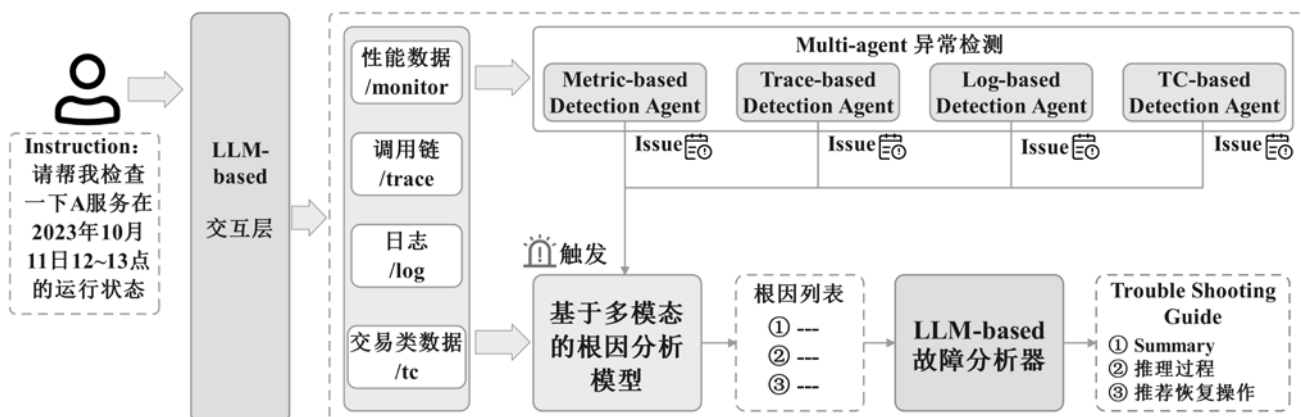
用标注好的历史云事件数据对模型进行训练, 让模型学习事件特征与归因之间的关联模式。

- » **归因推理:** 将新发生的云事件数据输入训练好的模型, 模型根据学习到的知识和模式, 对事件的可能归因进行预测和分析。
- » **结果解释与验证:** 对模型输出的归因结果进行解释和评估, 判断其合理性和准确性, 结合领域知识和人工经验, 对归因结果进行验证和修正。

- » **持续优化:** 根据新的数据和反馈, 不断调整模型的参数, 改进特征工程, 以提高归因分析的准确性和可靠性。

在整个过程中, 数据的质量和标注的准确性对模型的性能有着至关重要的影响。同时, 结合领域专家的知识经验和进行人工干预和验证, 能够有效提高归因分析的效果和可信度。

基于多智能体的多模态数据融合根因定位方法



以上整体架构方案包括四个部分：

- » LLM-based交互层：此模块用于理解用户的查询信息，提取出用户给出的基础任务和参数。使用ChatGLM2作为基础模型，在源文件代码中加入self-consistency、CoT和in-context learning的逻辑，使模型更能理解我们的场景并能更好地做出回答。
- » Multi-agent异常检测：由于涉及多种

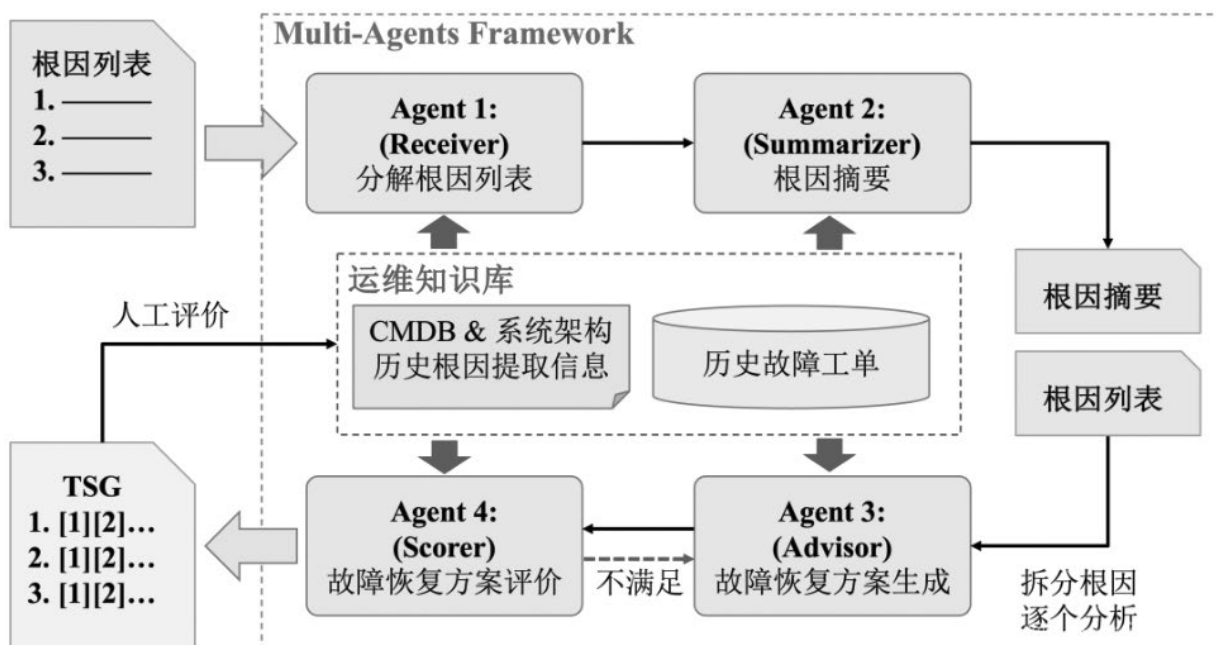
模态的数据源，采用单一检测模块难以获得高准确率且容易产生假阳性，因此我们采用multi-agent的检测方案。我们针对trace、log、metric三种模态数据均设计了异常检测agent。

- » 基于多模态数据融合的根本分析模型：算法将调用链、日志和指标等多模态数据转换成统一的事件表达，利用无监督的频繁项集挖掘的方法找出故障模式，在资源和代码块级别定位细粒度根

因。同时该方法还能通过对比故障前后的模式变化对故障进行解释。

- » LLM-based故障分析器：采用多LLM Agent轮询问答的方式，不同的LLM Session作为不同角色，生成故障报告工单。

在众多异常中，找到问题根本原因，故障报告生成能力。即根据分析结果生成故障报告和恢复建议。





我们同时开启四个LLM会话，作为四个具有不同角色的agent。第一个agent的角色是receiver，其接收上游输入的根本列表并拆解成多个自然语言描述的根本。第二个agent是summarizer，根据拆解得到的多个根本描述，汇聚形成根本摘要。第三个agent是advisor，对每个单独的根本进行分析，并给出对应的故障恢复方案。第四个agent是scorer，对方案进行评价，如果不满足预期，则重复进行新一轮迭代。

最后输出一个报告单，总结根本摘要和恢复方案。

AI Agents 自动运维报告单

故障问题清单

- 故障报告 ID: 5230c984-7e16-11ee-a836-965a262bc0e6
- 提交时间: 2023-11-18 15:45:39, 627260
- 问题描述: 对以下故障根本进行分析并给出恢复建议

疑似根本 1

- 故障问题: 接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧
- 排序: 1
- trace_id: trace_3a41df507be7a6157905a500b4532745
- 问题说明: 调用耗时 971ms 超过正常上限 561.41ms

疑似根本 2

- 故障问题: 调用链出现结构异常，或出现新的调用链模式
- 排序: 2
- trace_id: trace_6feb7529623b0e0f535a2be5ad55894
- 问题说明: 新增服务调用 span('Weblogic_19, Appld008, Sys001', 'Weblogic_19, Appld008, Sys001')

疑似根本 3

- 故障问题: 调用链出现结构异常，或出现新的调用链模式
- 排序: 3
- trace_id: trace_06f9aa5101e0u2657fadc9117e44a19
- 问题说明: 缺失服务调用 span('Weblogic_21, Appld008, Sys001', 'Weblogic_21, Appld008, Sys001')

的故障摘要

解析结果: 根据历史故障信息，这些故障同时发生时最大可能的原因是

- 接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧，可能是由于服务端处理能力不足或者网络延迟导致的。这可能导致调用链中的其他服务受到影响，从而出现结构异常或新的调用链模式。
- 调用链出现结构异常或新的调用链模式，可能是由于服务间的依赖关系发生变化或者某个服务出现了问题。新增的服务调用 span('Weblogic_19, Appld008, Sys001', 'Weblogic_19, Appld008, Sys001') 可能是为了解决某个问题而引入的新的逻辑。
- 调用链出现结构异常或新的调用链模式，缺失服务调用 span('Weblogic_21, Appld008, Sys001', 'Weblogic_21, Appld008, Sys001') 可能是由于

恢复方案:

- 故障恢复清单: 每个故障根本的恢复建议如下

疑似根本 1

- 故障问题: 调用链出现结构异常，或出现新的调用链模式
- 恢复方案:
 - 检查服务端的代码，确保逻辑正确。
 - 检查服务的配置文件，确保配置正确。
 - 如果可能的话，可以考虑使用调试工具或者日志来定位问题所在。

疑似根本 2

- 故障问题: 接口/coupon/txCtrl 的 rpc 方法 rpc 调用耗时加剧
- 恢复方案:
 - 检查网络延迟，确保客户端和服务端之间的通信畅通。
 - 优化服务端的代码，提高处理速度。
 - 如果可能的话，可以考虑使用负载均衡或者分布式架构来分散请求压力。

疑似根本 3

- 故障问题: 调用链出现结构异常，或出现新的调用链模式
- 恢复方案:
 - 检查服务端的代码，确保逻辑正确。
 - 检查服务的配置文件，确保配置正确。

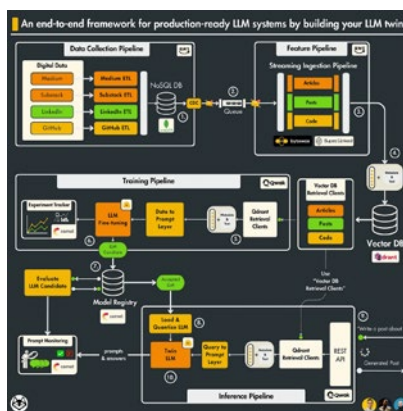
2023-11-19 14:41:23 Sunday November

大模型赋能智能运维

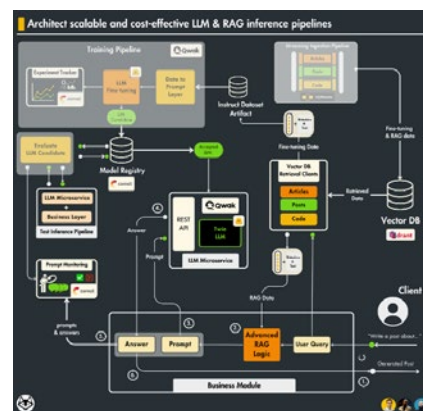
大模型可观测性

» 大模型训练和推理具有复杂的软件栈，亟需可观测性。大模型训练和推理的过程涉及到复杂的软件栈，这意味着在这个过程中，我们需要对各种组件和模块进行有效的管理和监控。这就需要一种可观测性的方法，以便我们理解和跟踪模型的行为，以及在出现问题时能够快速地进行调试和修复。

» 国外某创业公司给出观察大模型的5个维度即评价、调用链、提示工程、搜索和查询、微调实现大模型可观测性。大模型可观测性是一个复杂的过程，需要对模型进行全面的评价、追踪和提示验证。首先，我们需要对模型进行评价，以确定其性能和效果。然后，我

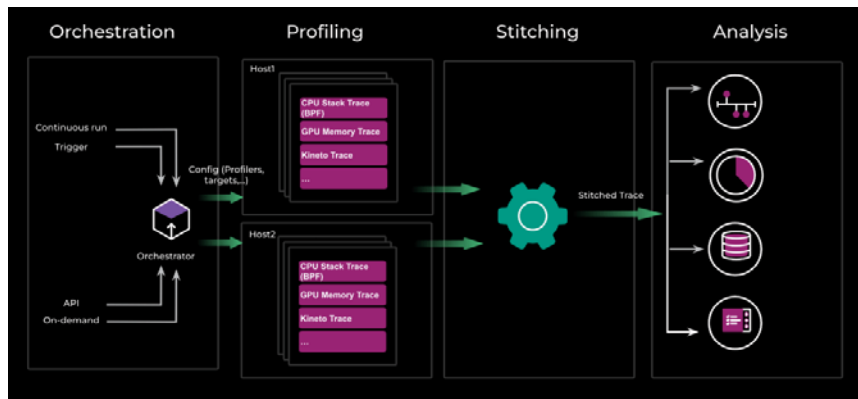


们需要追踪模型的运行情况，以便及时发现和解决问题状。接着，我们需要进行搜索和查询，以获取更多的信息和数据，以帮助我们更好地理解

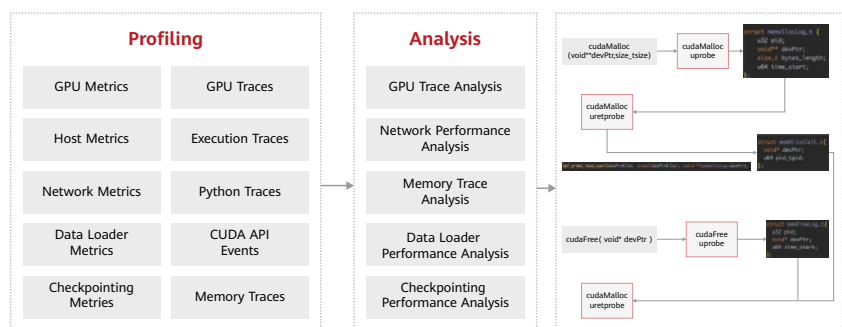


化模型。最后，我们需要进行微调，以实现大模型的可观测性，使其能够更好地适应不同的环境和需求。

» **Meta提出了基于profiling的多模态数据关联方法。**这种方法基于 profiling, 可以用于多模态数据的关联。多模态数据是指来自不同传感器或来源的数据, 例如图像、语音、文本等。这种方法的目标是将这些不同类型的数据关联起来, 以便进行更深入的分析 and 理解, 这种基于profiling的多模态数据关联方法具有很高的实用价值, 可以应用于多种领域, 例如医疗、金融、社交媒体等。



» **基于eBPF的跨层次、跨节点大模型请求追踪；**Meta提出了一种基于profiling的多模态数据关联方法, 该方法结合了基于eBPF uprobe的CUDA监控。通过对不同模态数据的分析和关联, 可以更准确地理解系统的运行状况和性能瓶颈。同时, 通过使用eBPF uprobe 技术对CUDA进行监控, 可以实时获取GPU的运行状态和性能指标, 为系统优化提供更精准的数据支持。这种方法在高性能计算和人工智能领域具有广泛的应用前景。

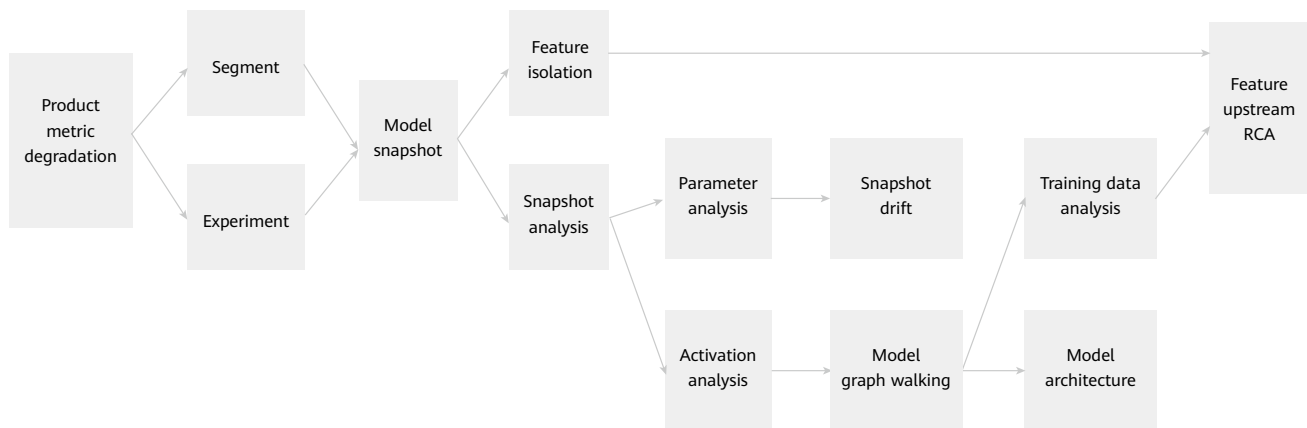


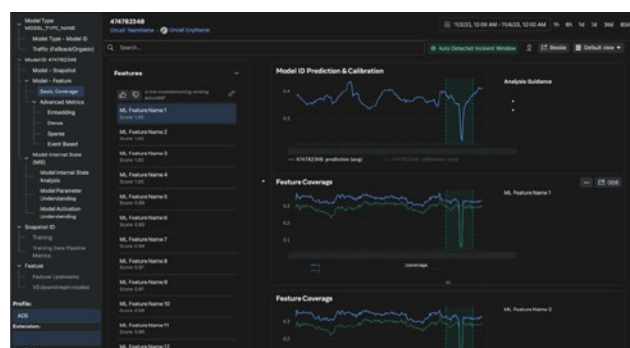
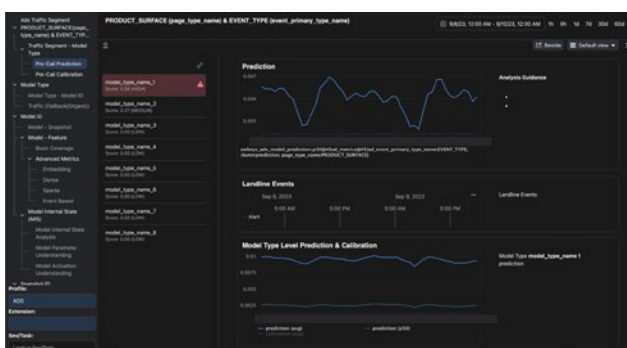
大模型故障定因

知识图谱驱动的渐进式根因定位方法。知识图谱驱动的渐进式根因定位方法是一种先进的技术, 它利用知识图谱的强大能力来帮助我们找到问题的根源。这

种方法的核心思想是, 我们可以通过逐步的推理和分析, 从知识图谱中获取到与问题相关的信息, 然后根据这些信息来定位问题的根源状因。这种方法的优势在于, 它可以帮助我们在复杂的问题中找到问题的根源, 而不是仅仅找到问题

的表面原因。这是因为, 知识图谱可以提供关于问题的全面和深入的理解, 包括问题的背景、相关的实体和事件、以及可能的原因和影响。通过这种方式, 我们可以更准确地找到问题的根源, 从而更有效地解决问题。



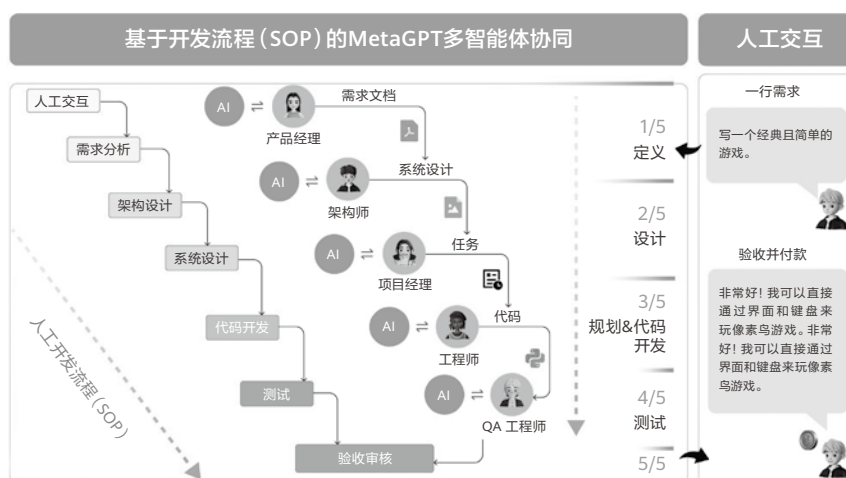


多智能体协作的智能运维。METAGPT 框架是一种非常强大的多AGENT协作框架，它可以帮助开发者更加高效构建多AGENT系统，并提供了多种不同的协作方式。在METAGPT框架中，开发者可以使用元编程技术来动态地生成AGENT代码，从而实现不同AGENT之间的协作。

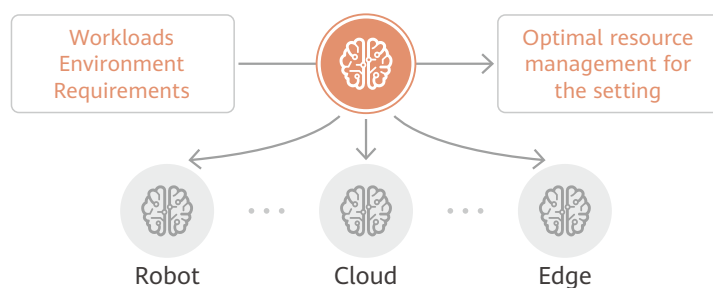
大模型驱动的操作系统。基础操作系统（base OS）是计算机系统的根基，它负责管理硬件资源和提供基本的系统功能。基于机器学习的端到端资源管理、工作负载环境要求、环境下的最优资源管理、机器人、云、边缘，基于单基座自动适应不同的设置。在现代计算机系统中，底层操作系统的设计和优化已经变得越来越复杂。为了提高系统的性能和效率，许多研究人员开始使用机器学习技术来进行端到端的资源管理。这种方法可以根据工作负载的环境要求，自动调整和优化系统资源的分配，以实现最优的资源管理。

基于确定性运维实践 实现大模型与智能运维双向奔赴

在大模型与智能运维的协同中，我们不仅仅使用LLM去搭建一个完整的端到端大模型，而是采用大小模型融合的思路。大模型主要负责处理一些不涉及具体系统本身，但需要按人类思维进行分析的任务，如决策制定、问题解决等。而小模型则专注于处理垂向问题的专业任务，如网络故障预测、性能优化等。这种大小模型的结合，可以充分利用大模型的通用知识，同时减轻LLM幻觉带来的危害。



ML-based end-to-end resource management



Single base auto-adapts to different settings

然而，大模型在应用的过程中，我们也面临着一些挑战。这些挑战都是在运维领域用好大模型需要去认真思考和回应的。我们需要找到一种方法，既能充分利用大模型的优势，又能有效地解决这些挑战。这可能需要我们在大模型的训练和使用过程中，不断地进行优化和调整，以适应不断变化的环境。

总的来说，大模型与智能运维的协同是一个复杂而富有挑战性的过程。我们需要在充分利用大模型的优势的同时，也要注意解决其带来的问题和挑战。基于确定性运维实践的经验，达成确定性的运维结果，只有这样，我们才能真正地将大模型应用于智能运维，为企业带来更大的价值。

AI大规模训练集群稳定性实战，加速智能跃迁



童琳
华为云SRE高级专家

摘要

本文通过 AI 大模型训练集群稳定性挑战分析，引出华为云针对 AI 集群构建的从异常感知、诊断到故障自愈的完整闭环能力及实战经验分享，为 AI 集群的稳定可靠提供有效的保障系统，加速智能跃迁。

”

AI大模型训练集群稳定性挑战

从行业看，AI大模型稳定性关注度越来越高，训练任务通信流向复杂、对网络时延和带宽敏感，AI大模型训练集群的可靠性面临如下挑战：

- » 伴随大模型参数数量的增长，集群规模越来越大，网络中元器件故障次数也会增加，整体通信网络更加容易受损，进而导致任务训练故障概率呈指数上升。
- » 大模型训练任务流量模型复杂，通信

链路长，通信跨域多类型的基础设施与云服务，故障类型包含带宽不足导致训练慢、损失发散、GPU Lost、训练任务运行失败、卡死、变慢等，原因和种类繁多，导致故障定界定位困难。

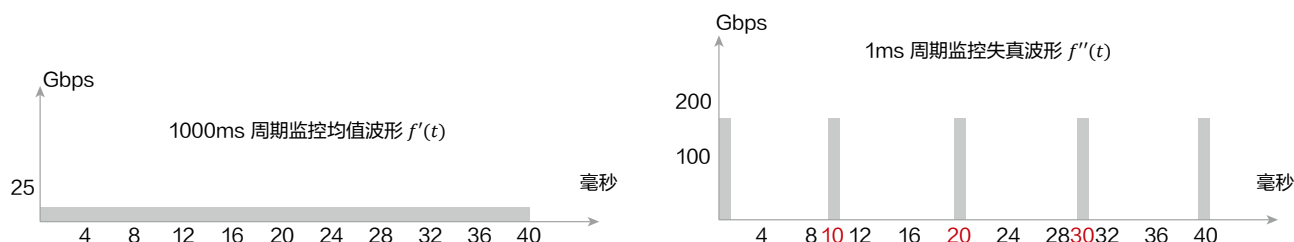
- » 故障发生后训练恢复的过程慢，特别是checkpoint加载非常耗时，一般长达几个小时，行业大规模集群MTBF仅数小时，每次故障时长长达10h以上。

- » 故障会导致训练任务不断回滚和重启，集群整体资源利用率低，浪费算力资源，业界AI训练集群平均资源利用率仅30%-40%左右。

综合来看，故障带来AI训练任务运行效率的问题和费用成本急剧上升。因此大模型集群的稳定、可靠将极大帮助企业提升效率、节约成本。

AI集群稳定性挑战一：监控精度不足，训练任务对网络依赖强

传统流量监控精度不足，难以准确监控流量特征



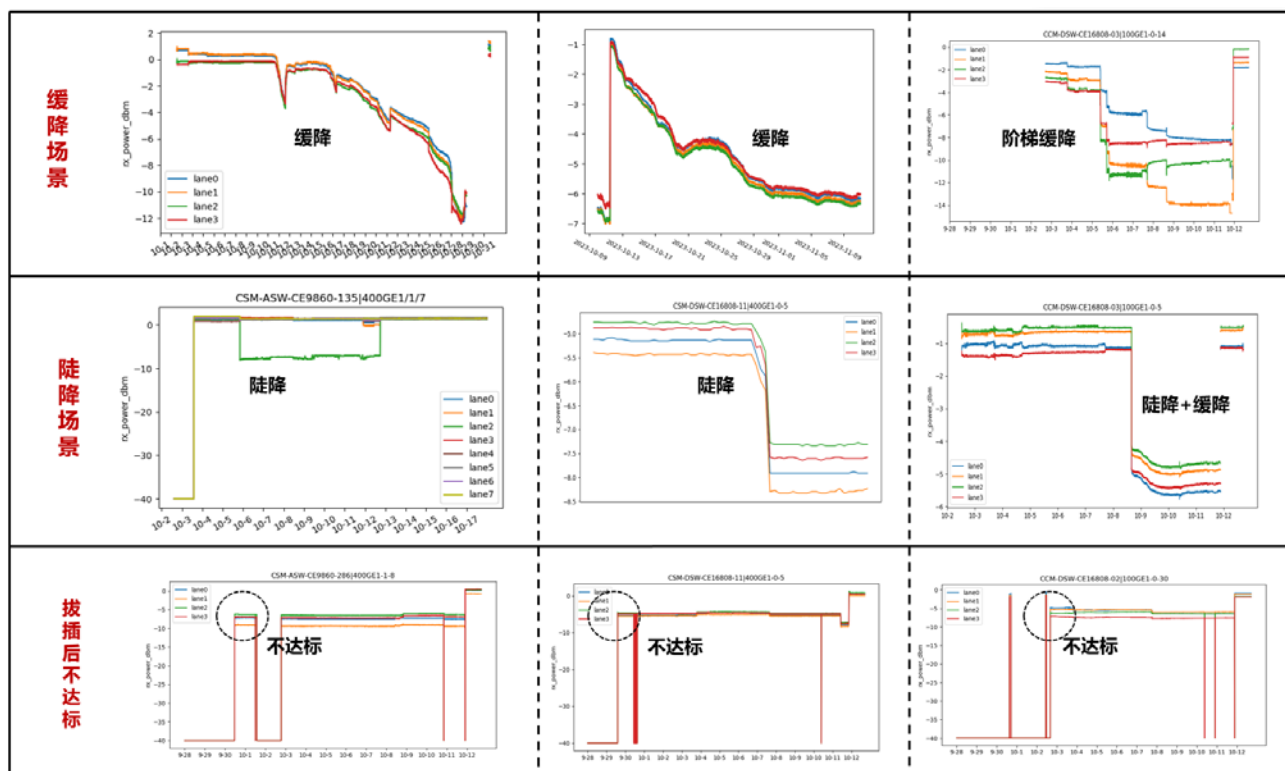
在AI集群的可观测性领域，面临采集精度不足、网络质量和性能检测要求高等问题。

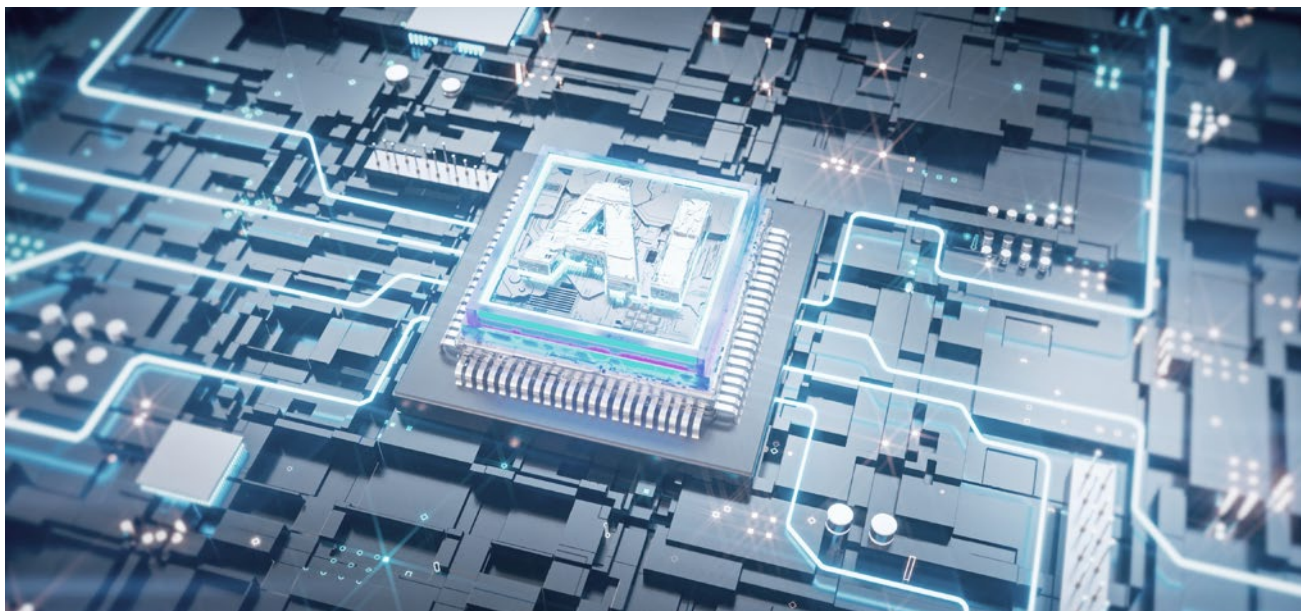
» 传统网络设备流量监控以SNMP协议为主，一般只能精确到秒级，而大模型训练以计算、通信两阶段交替进行，流

量波形变化以ms为单位来统计；另外传统监控采集计数器存在10~20ms的软件误差，导致监控结果失真和错位，最终结果是无法精确测量AI集群训练任务的真实流量特征；

» 传统运维只能检测光模块彻底故障，

由于光模块长期运行，性能逐步衰减，形成亚健康状态，且故障特征多种多样。伴随AI集群规模增加，光模块亚健康对AI集群训练任务影响会逐步放大，因此光模块故障亚健康检测、以及故障的预防在AI训练集群中比传统应用集群更加重要。





AI集群稳定性挑战二：快速定位定界难度大，故障恢复慢

在业务受损场景下，特别是硬件和网络亚健康状态，AI集群故障的定界定位极其困难：

» 故障模式复杂多样：慢节点、慢网络、慢计算、慢通信等疑难故障，均体现在训练业务中断、卡顿、失败，而某些类

型故障无明显白盒指标告警，一般难以快速定界；

» 故障链路长：大规模训练任务涉及多对多的GPU通信，且通信链路涉及的服务器、设备、端口、链路、光模块，跨越多种网元，业务故障需关联分析的

对象种类多；

» 诊断效率低：传统诊断依赖训练任务的海量日志，分析时间长，同时由于缺失网络流量实时拓扑，需要分析海量监控指标和数据。

故障模式多

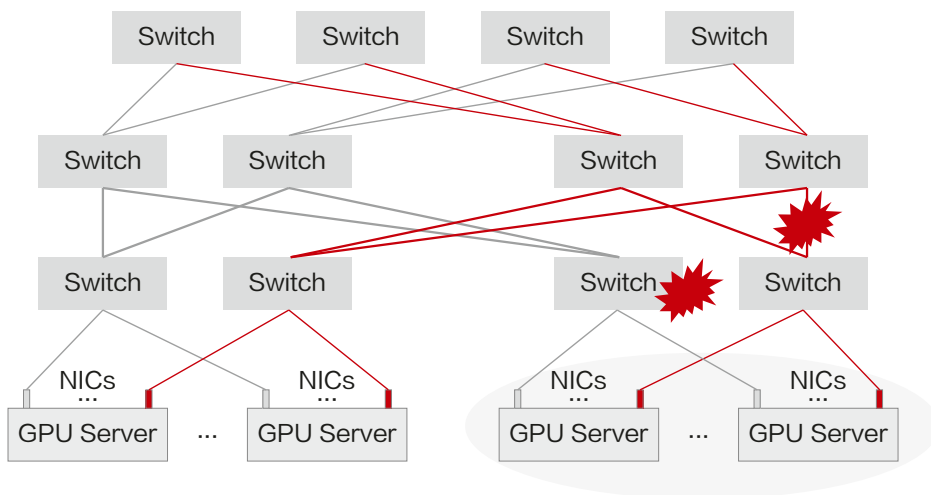
慢节点、慢网络、光模块性能劣化

故障链路长

通信链路设计服务器、设备、光模块

诊断效率低

面对海量的日志和监控指标



AI集群稳定性平台能力构建

为了应对上述挑战，华为云在内部成熟运维平台基础上，针对AI集群构建了一套从异常感知、诊断到故障自愈的完整闭环能力，构建AI集群稳定性平台能力全景图。

» 故障感知阶段，通过故障快速检测和预测，尽可能先于客户发现故障，在华为内部已构筑内存/硬盘/光模块等对

象的故障预测算法，能提前发现大部分硬件隐患和亚健康故障；同时感知系统将持续自动采集网络连接拓扑和流量路径变化，为诊断提供全面的数据支撑。

» 故障诊断阶段，通过全链路拓扑还原技术，自动绘制业务实时拓扑，结合流特征实时的差异性变化，能快速识别

网络拥塞问题，同时结合关联对象的指标、日志、告警，快速进行故障根因溯源和推荐。

» 故障恢复阶段，决策系统基于图配置库，实时分析计算故障影响面（包含影响的节点范围、客户等），基于设定好的故障预案模型，使用代价最小、效率最高的策略来快速恢复业务。

1min感知，5min诊断，10min自愈



白盒和黑盒相结合的可观测性能力，监控无盲点

依托于华为云大数据的平台能力，构建毫秒级指标采集、存储、计算、分析的监控平台，支撑AI集群的可观测性能力落地。在传统监控系统基础上，强化以下3个关键能力：

» 高精度指标采集能力：通过Telemetry采集技术，除针对传统的接口速率、报文丢弃的毫秒级监控之外，还实现了

PFC收发报文、队列拥塞丢包、端口带宽使用率等关键指标的毫秒级感知能力，能快速感知网络微突发场景。

» 网络丢包实时感知能力：通过网络芯片内置丢包检测能力，主动检测流级丢包异常和原因，通过分布式探针采集和压缩、汇聚，上报到服务端形成事件并进一步传输给监控平台进行

告警。

» 黑盒监控能力：不同于传统的pingmesh是基于ICMP、TCP、UDP等标准协议来拨测，针对AI集群的隐形丢包，构建基于HCCL通信协议集合的链路级全覆盖的黑盒拨测系统，慢网络等故障可秒级感知。

AI加持的光模块亚健康预测, 提前预防问题发生

针对光模块故障检测和预测领域, 通过空间、时间和多维指标来提升故障检测能力。具体来讲基于光模块的历史故障特征(厂商、型号、故障原因等)、日志特征、指标特征(电流电压、温度、光功率等)、时间维度(使用年限)等综合因素,

结合光模块拓扑链路本端和对端的相似指标综合分析。

结合这些多维度的数据, 利用特征工程、回归分析AI算法来识别已发生故障、和正在发生指标劣化但尚未发生

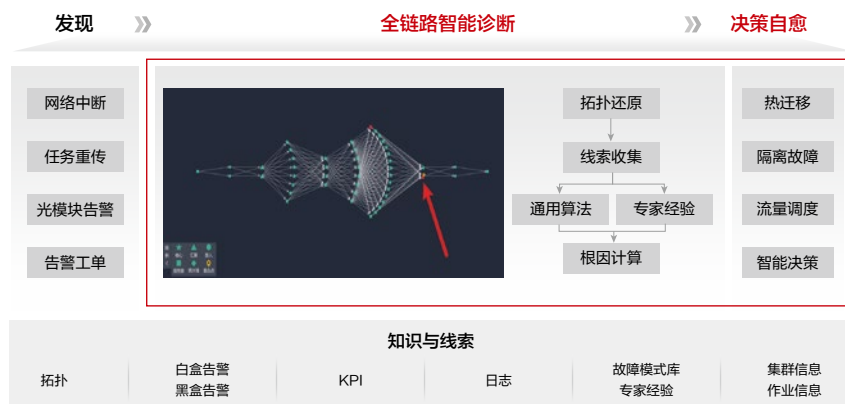
故障的光模块(包括脏污、松动等故障场景), 并进行故障画像评分和寿命预测, 生成检测报告, 并按照阈值自动生成告警和提交维修单, 这样就可以在业务受影响之前, 替换掉将会发生故障的光模块。

基于全链路路径还原技术, 5min内直达问题根因

针对AI集群的故障快速定界, 通过自研的全链路诊断系统, 结合全景图配置数据库, 能快速计算训练任务涉及的任意1台源和1台目的主机的流量路径, 路径涵盖服务器、GPU卡、端口、链路、单板、交换机等所有网元。

诊断分析模块结合实时拓扑中对象的指标、日志、告警等, 通过知识图谱进行故障根因推荐, 在数分钟内识别出故障单元, 该技术目前广泛使用在华为云内部, 确保了故障定界时长<5min。

决策恢复模块具备全自动化故障预案处理能力, 支持端口故障隔离、流量调度、任务迁移等手段, 快速恢复业务。



以上是我们针对AI大模型训练集群稳定性挑战的一个完整实践, 通过白盒监控和黑盒监控相结合进行故障秒级感知和预防, 再基于全链路拓扑故障诊断到具

体网元, 最后通过三板斧进行故障恢复, 从而为AI集群的稳定可靠提供一个有效的保障系统, 这样AI的业务发展才能加速跃迁。



重塑千行万业，加速智能化转型

基于AI原生应用引擎构建行业解决方案，一站式解决企业AI场景落地问题，
助力企业智能化转型，实现创新增长、优化决策、降本增效！



垂直行业应用场景 100+ 个，通用应用场景 50+ 个

智能报表

智能客服

辅助研发

知识问答

辅助办公

智能会议

人力资源

工程实施

大模型时代下的应用可观测性方案探索与实践

— 华为GTS与基调听云在面向未来的应用可观测性方案上的探索与实践

作者: 王福强 杨金全

摘要

本文主要介绍华为 GTS 与基调听云基于大模型技术的应用可观测性方案的探索与实践，通过分析其背景、实施方案以及落地效果，以创新技术提升应用的可观测性，确保企业能够在复杂多变的技术环境中保持竞争力。

王福强

华为云云服务SRE专家，深耕SRE运维领域，曾主导180+租户快速从其他云上搬迁至华为云，构建云服务端到端的运维流程及规范体系，具有丰富的运维经验。

杨金全

基调听云首席技术官（CTO），在智能观测性平台领域拥有丰富的研发和商业化经验，作为早期商业化APM（应用性能管理）工具的开创者，在此领域具有深刻见解。

背景

在当今数字化转型的浪潮中，企业越来越依赖复杂的应用程序和服务来支持其业务运作。这些应用程序不仅需要快速响应，还必须具有高可用性和可靠性。为了确保这些应用的性能和稳定性，应用性能管理（APM）和可观测性工具变的至关重要。

传统的APM工具主要集中在监控应用的性能指标，如响应时间、吞吐量和错误率等。然而，随着技术的快速发展和业务需求的变化，单纯的性能监控已经无法满足现代应用的需求。云计算、微服务架构、容器化技术、DevOps实践、人工智能、物联网（IoT）与边缘计算、5G技术的普及，共同推动了应用可观测性方案的

创新和发展。这些技术不仅为企业提供了灵活的资源管理和服务部署能力，也带来了新的监控挑战。企业必须构建更为先进和全面的监控系统来应对云环境的动态性、微服务的复杂性、容器的短暂性和DevOps的快速迭代。同时，人工智能和机器学习的进步为监控数据的深度分析提供了新工具，而5G和IoT的发展则进一步扩展了监控的范围和深度。

现代应用不仅需要监控性能指标，还需要具备更强的可观测性，包括日志、指标、分布式追踪等数据源的整合与分析。仅依赖传统的监控方法已无法全面反映应用程序的健康状态和运行性能。



在 GTS 内部, 所有应用均已改造为依托于华为云, 基于云原生技术构建的现代应用。为保障这些应用的稳定性, 历史上已经建设了非常多的系统或工具, 建立了一套一体化监控+AIOPS 的运维监控体系, 在当前看来存在一些明显问题。



那些年我们一起趟过的坑

基于大模型技术的可观测性方案



大模型时代的可观测性全景

在大模型时代，可观测性平台也需要与时俱进，利用大模型技术，可以实现更智能化的故障检测和预警，自动化的根因分析以及精准的容量预测。利用大模型对海量数据的分析处理，整合及关联不同数据源，可观测性平台不仅能够提高运维效率，还能提供更准确和及时的决策支持，全面提升系统的稳定性和可靠性。

基于以上全景图，GTS 和基调听云提出大模型时代的可观测性方案主要包括以下几个方面：

1 建立可观测性数据模型

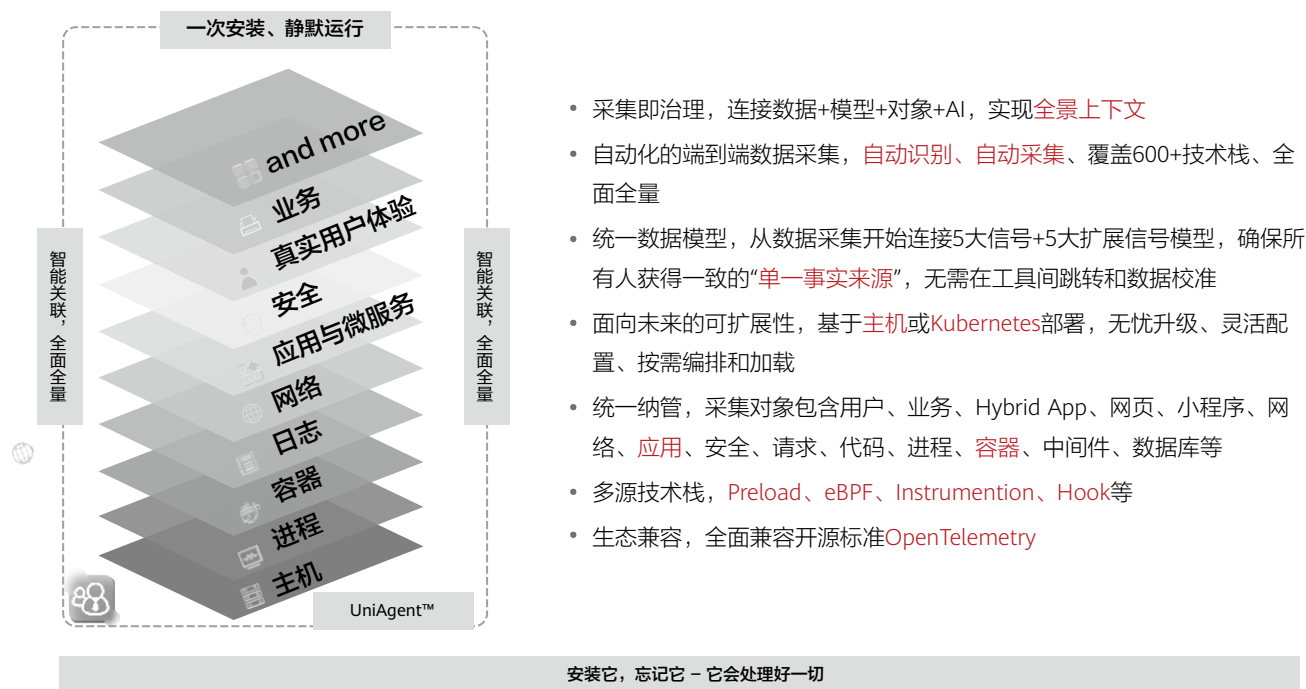
建立系统化的可观测性数据模型是提升

应用可观测性的关键。该模型包括几个核心要素：首先，模型分类在CNCF的五大信号基础上，新增了拓扑、业务属性、用户旅途、元数据和网络等信号，形成了十大全面的数据模型。数据源定义则明确了日志、指标、追踪数据等各种数据源的类型和结构，包括传统监控数据和新兴的大模型生成数据。通过数据分类与标签化，对数据进行系统分类和标签，便于后续的查询和分析。数据关联建立了不同数据源之间的关系，提供了全面的数据视图，支持跨模型、跨系统和跨应用的深入分析。数据标准化制定了统一的数据格式和标准，确保不同数据源的数据能够无缝集成和分析。这一综合模型将显著提升应用的可观测性，帮助企业

更高效地利用数据，优化系统性能。

2 运维对象的自动纳管和全面全量的数据收集与融合

自动纳管通过自动化工具和智能平台，实现对服务器、应用、服务等运维对象的自动发现和纳管，确保新部署的资源自动纳入监控范围。全面全量数据收集使用分布式技术，保证日志、指标、分布式追踪等所有关键数据的完整性和准确性。数据融合则将不同来源的数据整合，形成统一的数据视图，实现跨系统的数据集成和分析，提供更全面的可观测性。这些技术共同提升了运维的自动化和智能化水平，为系统的高效管理奠定基础。



UniAgent 采集即治理

采集即治理是本方案的核心，通过对数据模型、数据计算标准和数据关联关系的治理，为 AI 提供高质量可观测性数据。



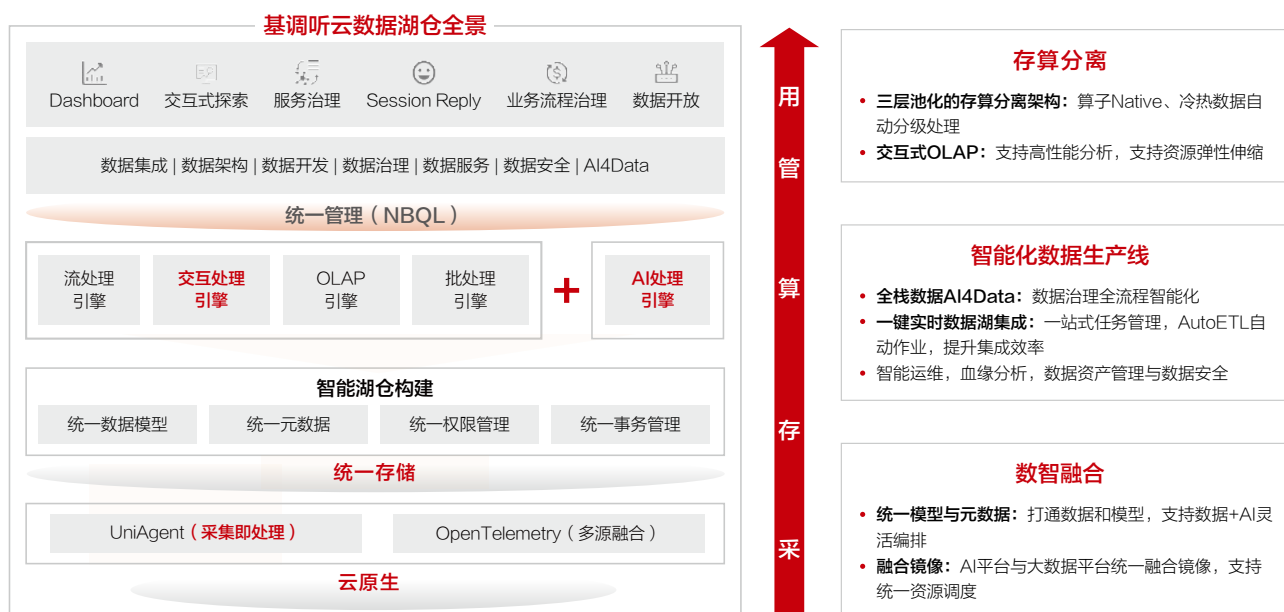
3 可观测性基础软件平台建设

该平台具备高可用、高性能的数据处理和存储能力，以支持大规模数据的实时处理和分析。在面向未来的可观测性体系中，数据湖仓是可观测平台的核心技术。

数据湖仓技术将数据湖与数据仓库的优势结合，提供了一个统一的数据存储和

管理平台。数据湖支持存储和处理各种类型的原始数据，而数据仓库则优化了结构化数据的查询和分析。通过融合这两者，企业能够在一个平台上实现数据的存储、管理和分析，提升数据处理的效率与灵活性。数据湖仓技术利用分布式存储系统和列式存储系统来支持海量数据的存储与高效查询，同时通过元数据管理和数据治理工具保障数据质量与一

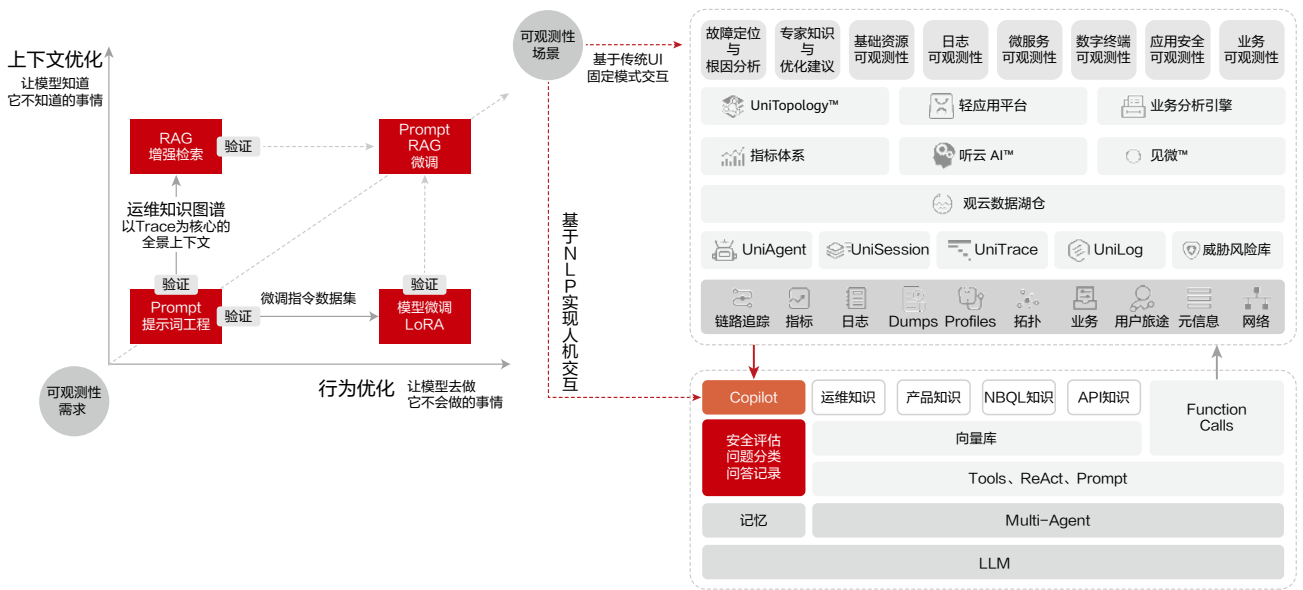
致性。在数据处理与分析层面，数据湖仓平台结合大数据处理框架和交互式查询工具，实现数据的批处理、流处理、分析与可视化，其良好的可扩展性和高可用性通过分布式架构和多副本存储机制确保系统稳定运行，满足不断增长的数据需求。这种技术已成为构建现代化可观测性平台的核心，助力企业在大数据时代高效决策和运营。



数据湖仓：全栈可观测性大数据基础设施

4 大模型应用建设

大模型技术为可观测性引入了智能分析的新维度。通过利用大数据平台和机器学习框架，大模型在异常检测、根因分析和预测分析中得到了优化。它能自动分析日志、指标和追踪数据，识别异常模式、定位故障根因并提供优化建议。同时，大模型还支持多源数据融合与深层次信息挖掘，帮助发现潜在问题和优化机会。



基于大模型的可观测性平台架构

5 统一运维体系建设

在未来的应用可观测性方案中，统一运维体系建设是确保系统高效运转和跨部门协作的关键。该体系包含四个核心能力：Dashboard、指标体系、多维分析和全局拓扑。

» Dashboard提供了一个集中化、可视化的界面，用于实时监控系统的健康状况

态和性能，并支持自定义视图和动态交互分析。

» 指标体系定义了需要重点监控的各种系统性能和业务指标，包括关键性能指标 (KPI) 和告警阈值设定，为系统稳定性提供衡量标准。

» 多维分析则通过多维度数据切片、趋势分析和根因分析，深入探讨系

统行为和性能变化，帮助定位问题根源。

» 全局拓扑可视化展示系统中各组件和服务的关系，支持动态拓扑视图和故障传播路径分析，实现依赖关系管理和自动化拓扑更新。

这一体系为企业提供了全面的运维视角，确保系统的高效、稳定运行。

6 智能运维 (AIOPS) 体系建设

GTS和听云在智能运维体系的建设中, 重点关注异常检测、告警收敛、基于因果关系的确定性根因分析、故障管理和容量管理等关键模块, 以打造一个更智能、更高效的运维生态系统。

在智能运维 (AIOPS) 体系中, 异常检测、告警收敛、基于因果关系的确定性根因分析和故障管理是核心模块, 确保系统的稳定性和高效性。

» 异常检测通过多维数据分析和机器学习模型, 实时识别系统中的异常行为, 自动调整阈值以适应动态环境, 并在检测到异常时立即触发告警。

» 告警收敛则通过告警聚合、优先级排序和智能关联分析, 减少告警噪音, 确保运维人员专注于关键问题。

» 基于因果关系的确定性根因分析利用因果关系模型和历史数据匹配, 快速

定位故障根源, 并提供实时更新的分析报告, 提升根因分析的精准度。

» 故障管理模块通过自动化手段检测和修复故障, 详细记录并追踪每次故障的全过程, 推动系统的持续改进与学习。

这些模块的集成为企业构建了一个智能、高效的运维体系, 显著提升了系统的可靠性和响应速度。



总结与展望

在数字化转型的背景下, 企业应用系统变得越来越复杂, 这对传统的运维方式提出了巨大的挑战。GTS和听云通过引入智能运维 (AIOPS) 体系, 构建了面向未来的可观测性方案, 利用先进的人工智能和自动化技术, 不仅提升了运维效率, 还提高了系统的稳定性和可靠性。

智能运维体系的实施让企业能够更快速、更精准地响应系统问题, 减少业务中断, 并优化资源利用。通过大规模数据分

析和机器学习模型的应用, AIOPS体系能够自动识别问题、优化资源配置, 并在复杂环境中保持高效运行, 为企业带来了巨大的运营价值。

展望未来, 随着大模型技术和人工智能的进一步发展, 智能运维将会变得更加智能和自动化。未来的AIOPS体系将不仅仅局限于被动的监控和响应, 而是能够主动预测和预防潜在问题, 实现真正意义上的自愈和自优化。同时, 随着5G、物联网

和边缘计算等新兴技术的普及, 运维的可观测性将进一步扩展, 涵盖更多的应用场景和数据源, 为企业提供更加全面和深入的运维支持。未来的运维体系也将朝着更高的智能化、自动化和自适应性方向发展, 为企业的数字化转型提供坚实的技术保障, 并推动业务的持续创新与发展。在这个过程中, 华为GTS和听云将继续引领行业前沿, 探索并实践更加先进的可观测性方案, 助力企业实现卓越的运营效率和业务增长。



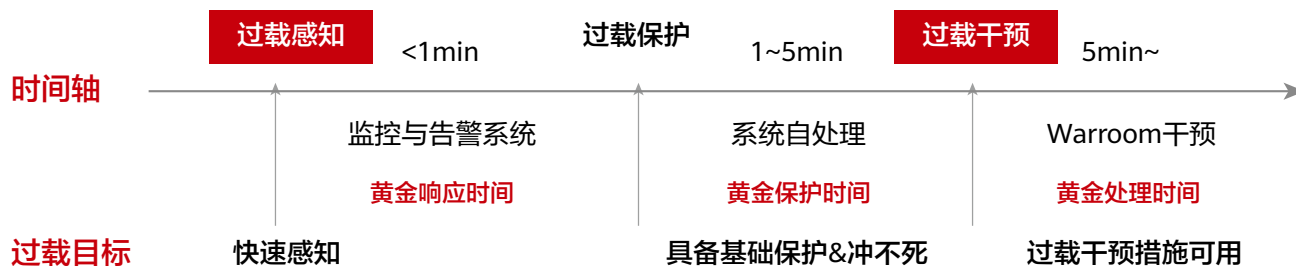
云服务过载控制“三部曲”

作者: 郑磊

摘要

云服务场景下面临不确定的客户应用，不确定的公网请求变化，不确定的系统负载挑战，如何做到云服务、微服务及租户的高可用，避免这些不确定的因素产生服务过载，如何使租户面的业务不受影响是本文重点讨论的话题。

过载控制有非常重要的三部曲：过载感知、过载保护、过载干预，其关键能力需要围绕这三部曲进行规划与建设，本文重点介绍过载感知与过载干预的实践能力。



过载现象

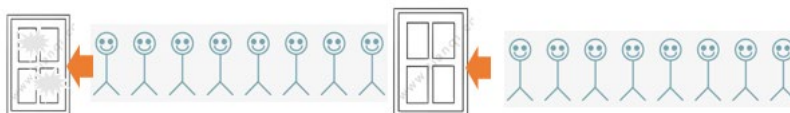
现象1: 关键节点出现过载, 所有的请求都出现异常。典型案例: 如黑客发起对网络服务超百倍的瞬间查询, 结果导致数据库CPU冲高, 其他正常客户的网络请求被阻塞。



现象2: 无序抢占, 每个请求/每个租户都出现性能损失。典型案例: 某个租户发起对所使用存储池的未设置有效的QoS控制, 造成瞬间XXXGb/s的带宽请求, 所有使用该存储池的用户延迟增加, 出现过载。



现象3: 性能没问题, 但是依赖的服务出现性能瓶颈。典型案例: 几个个VM同时请求云解析服务, 超过了安全封堵系统的边界阈值, 造成云解析服务处理正常, 被安全封堵系统禁止。



过载感知

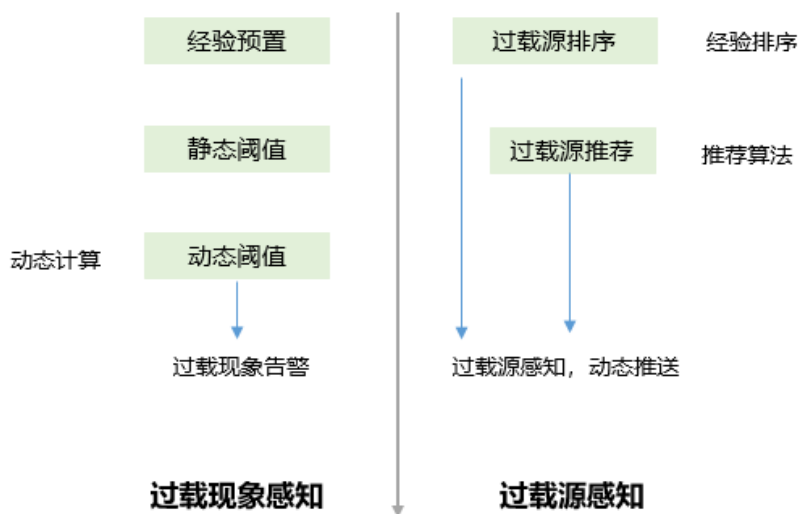
1 过载感知的关键能力

过载感知是过载控制的基础能力, 需要尽可能早的感知到过载现象和过载源, 进而支撑过载对象事件的响应和对过载源的及时处置, 达成分钟级感知过载现象和过载源的目标。

过载现象与过载源感知能力建设阶段

过载现象感知能力: 基于监控告警服务在线获取到过载现象。主要从固定阈值(经验预置), 和动态阈值(算法推荐)方式获取。

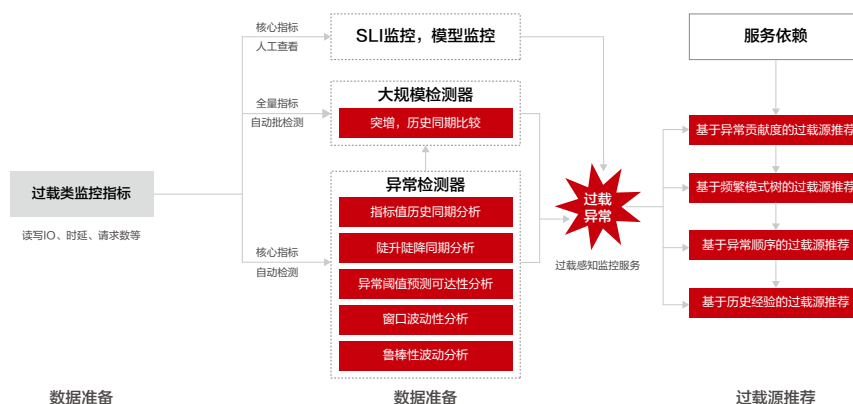
过载源感知能力: 基于监控告警服务在线获取造成过载的源头, 即谁的什么行为造成了过载。包括两种方式: 第一种是基于已知经验的计算和排序; 第二种是基于算法的在线推荐, 给出可能的过载源, 支撑过载源的在线感知、在线推荐和在线对象关联。





2 过载感知的能力架构

建立完整的“业务监控模型”，监控整个业务流程的健康状况，无需过多手工筛选和多个系统切换，准确发现、快速定位问题，实现分钟级感知过载现象，获取过载源。设计基于工作负载数据分析的过载感知算法，支撑过载现象感知、过载源快速准确推荐目标达成。



3 过载感知实践

解决异常度高但非根因的误诊断场景：



减少负相关指标对异常排序的干扰场景：



过载干预

过载干预主要是通过对服务的状态感知, 并通过在入口层、能力层、服务层和数据层进行分层控制与治理。通过感知过载及时获取过载现象, 通过过载控制服务主动采取过载控制措施, 达成服务的平稳运行, 控制影响半径和范围。

过载干预关键能力:

1 入口层

支持基于API接口、租户等级、租户名、URL、域名粒度的过载控制。

2 能力层

支持过载时的自动扩缩容, 支持基于接口/用户/域名的限流, 支持基于接口/用户/域名的降级, 支持按照用户/接口的熔断, 支持按照集群/按节点/按微服务的隔离。

3 服务层

支持集群级的过载控制、本地节点/服务的过载控制, 基于动态策略的过载控制。

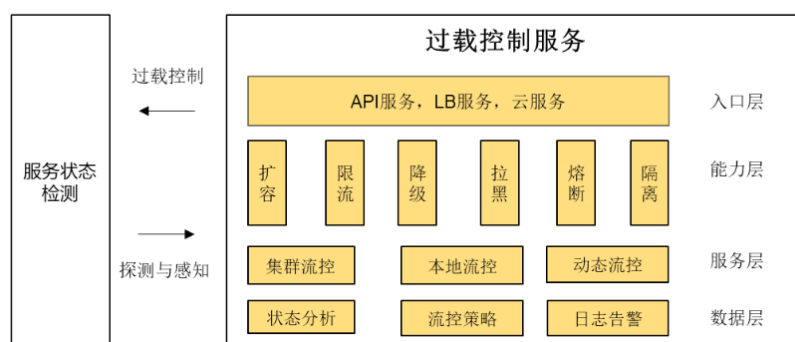
4 数据层

- » 流控策略: 支持策略集中下发以及基于作业优先级的过载控制, 支持基于作业时延的过载控制, 支持基于作业负载的过载控制。
- » 状态分析: 支持过载控制结果分析, 过载控制的时效性, 过载控制的假阳性、

假阴性结果分析。

- » 日志告警: 实现与过载感知信息的互动, 支持过载日志记录, 相关过载控制有监控告警的基础能力, 支持过载感知, 及过载控制的时效性感知, 过载控制收敛与治理效果监控可视。

过载控制的典型架构



过载控制的典型架构

结束语

在此, 引用都江堰的故事为过载控制做更通俗易懂的解释。古时, 四川盆地水患问题长期困扰当地居民且不可控, 古人利用西北高、东南低的地理条件, 根据江河出口处特殊的地形、水脉、水势, 顺势利导, 无坝引水, 自流灌溉, 实现了都江堰的水患的有效治理。关键能力有:

- » “鱼嘴”, 承担都江堰的引水工程。正常雨季保证了当地人们的农田灌溉和饮水, 雨季避免发生洪水。

- » “飞沙堰”, 承担都江堰的防洪作用。遇洪水时自行冲破堤岸, 让大量河水流入外河, 同时形成漩涡, 把多余的沙石抛到外河, 定期清洁。
- » “宝瓶口”, 承担都江堰的“控制门”作用, 整体比较窄, 有限制水流的功能。

云服务开发和使用场景下, 客户的一次业务促销, 公网的一次网络风暴, 外部的一

次黑客批量攻击或安全扫描, 这些都会造成云服务请求过载, 或者作业过载排队。本文介绍的过载感知、过载干预的实践有效支撑了云服务及云上应用过载类问题的有序分析, 系统性治理。希望这些实践和能力可以为云服务开发及应用开发的过载控制挑战带来参考, 也欢迎感兴趣的同仁们参与到过载控制类技术难题分享和讨论, 探讨新技术与关键能力的演进。

政务云架构设计与应急切换能力优化 实现业务安全稳定高质量

作者：康镇 李国强 刘纯纯 殷阁朕

摘要

本文主要介绍政务云 SRE 基于华为确定性运维能力体系，通过一系列系统高可用架构优化以及应急演练能力优化方案和用云优化实践，实现业务安全稳定高质量，助推政务客户实现数字化转型。

背景

国家数据局会同有关部门对外发布《关于深化智慧城市发展 推进城市全域数字化转型的指导意见》明确了推进城市全域数字化转型的发展方向，同时对城市重点领域提出了要求，包括基础设施建设、数字

经济培育、城市精准治理、城市环境优化与保障城市安全等方面。这意味着政务云在提升服务效能、创新治理模式的同时，更要兼顾云上业务的安全可靠、资源高效与业务敏捷。在《华为云确定性运维专刊

第四期》中，政务云SRE团队分享了政务云维护体系中的可观测性、安全性、容量健康性提升优化方法，本期将继续结合政务云架构设计可靠性与应急切换能力的优化方法与优秀实践，与大家进行分享探讨。



政务云业务面临的挑战分析

从早期的“政府信息化”阶段到“电子政务阶段”，再到以政务云作为底座的“数字政府”，随着政府运行、民生服务、社会治理和经济发展等各领域纷纷部署在政务云上，数据价值凸显，威胁攻击越来越多，尤其近年来，全球出现过多起如关键系统停机、云被黑导致数据丢失、云故障导致网站无法访问等质量事件，政务云业务连续性和安全挑战越来越严峻。具体到城市的政务云建设中，挑战主要来自以下几个方面：

一是重建设轻运营：常规的政务云更注重于云上基础设施建设，缺乏针对政务云平台稳定可靠运维体系和安全合规管理能力；同时，人工智能等新兴技术与运维融合深度不足，在业务稳定性面临新场景下的“适数化”要求下，传统运维方式存在可用性管理薄弱、被动响应、运维流程冗杂、故障管理能力欠缺、运维工具和业务需求不匹配、运维效能低等一系列问题，

都在制约着数字政府的深化发展。

二是重发展轻韧性：随着全域数字化转型深入发展，政务云系统越来越复杂，数据孤岛增加，版本迭代经常导致回退。如何避免系统出现故障，故障出现后如何快速恢复业务，成为亟需提升的能力。

面对以上挑战，政务云SRE基于华为确定性运维能力体系及解决方案，通过一系列政务云用云优化实践，实现系统高可用，适配架构改进，快速发现系统隐患，帮助政务客户完善运维体系，提升运维效能，助力政务客户实现数字化转型。

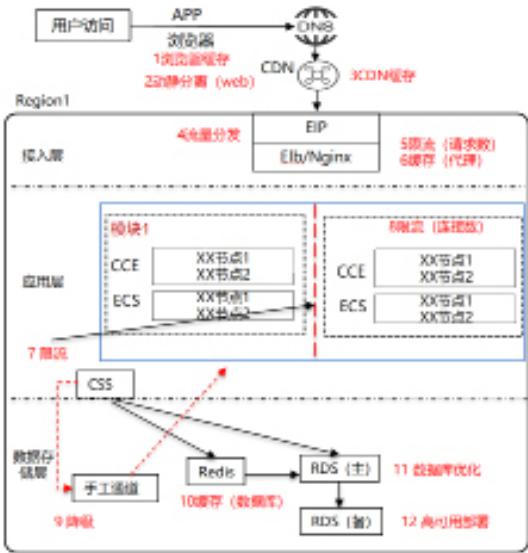
1 架构设计可靠性优化

我们通过调研业务架构、应用及技术架构，聚焦高可用架构能力建设，识别架构及业务系统部署的潜在风险，总结出针对不同业务场景的优化建议：

- » 应用架构优化：针对各业务模块或者子系统性能及扩展性问题，提供动静分离、流量分发、隔离解耦、限流、降级等场景应用架构优化方案。
- » 数据优化：为提高访问数据速度，结合业务访问量及数据流向，针对业务特性提供缓存及数据库优化等场景方案来优化数据。
- » 高可用部署优化：为解决云资源单点、数据无备份等问题，同时为了最大化的提升业务稳定性和可用性，政务云SRE团队以最优的业务资源开销协助客户进行高可用部署改造来优化高可用部署架构。

以某政务网站为例，从某网站架构的接入层、应用层、数据存储层进行全链路分析，展示各优化场景及其价值。

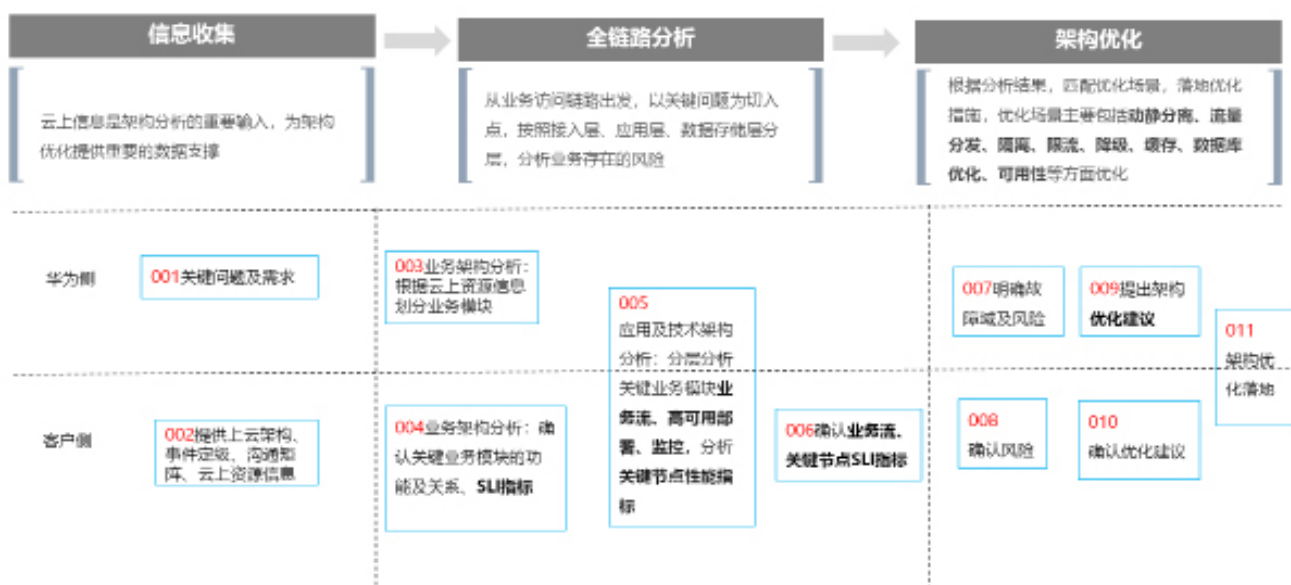
序号	架构分层	优化场景	价值
1	接入层	浏览器缓存	高性能-提高网站的访问速度
2		动静分离	高性能-页面加载加快
3		CDN缓存	高性能-降低访问时延，减轻源站主机压力
4		流量分发	可扩展-弹性伸缩，提高系统的处理能力
5		限流	高性能-避免大量请求击垮系统
6		代理缓存	高性能-提升访问速度，减轻源站服务器的压力
7	应用层	隔离	可扩展/高可用-模块化、组件化，服务故障不影响其他服务
8		限流	高性能-避免大量请求击垮系统
9	数据存储层	冗余	高可用-有损服务提升关键服务可用性
10		缓存 (Redis等)	高性能-缓解数据库负载，提升吞吐量
11		数据库优化	高性能-慢SQL优化，提升读写效率
12	可用性部署		高可用-提高关键节点可用性





整体业务分析优化流程分为三大部分：

- » 信息收集：云上信息是架构分析的重要输入，为架构优化提供重要的数据支撑。主要包括关键问题及需求，收集整理业务上云或云上阶段关键需求及问题，用于辅助识别业务架构风险。同时，收集相应的沟通矩阵，应用的事件定级标准、云上的资源及当前云上架构。
- » 全链路分析：从业务访问链路出发，以关键问题为切入点，按照接入层、应用层、数据存储层分层，分析业务存在的风险。以信息收集为输入，模块化分析业务架构，明确业务模块功能、关键SLI指标、业务高峰时间、流量变化（是否突降突增）、强相关模块、事件定级等关键信息。
- » 架构优化：分析现网指标、关键SLI，识别架构存在故障域及风险，匹配优化场景，确认优化措施并落地，优化场景主要包括动静分离、流量分发、隔离、限流、降级、缓存、数据库优化、可用性等方面优化。



2 应急演练能力优化

政务云应急演练能力优化服务遵循混沌工程实验原理，并融合华为云内部SRE团队“朱日和”突击演练实践，聚焦多领域模拟真实环境，提供丰富的演练场景，以及关键场景应急预案，进而提升政务客户对云上业务系统稳定性的信心。

应急演练-预期收益如下：

- » 发现运维人员协作沟通和运维技能方面存在的不足；
- » 发现应急预案中存在的问题，检验应急预案的可行性以及应急反应的准备情况；
- » 建立对系统抵御生产环境中失控条件的能力及信心；
- » 验证云上业务高可用方案的有效性（冗余保护、数据保护、容灾保护）。

整体演练方案

• 演练规划

制定切换演练规划，收到客户切换演练需求后，由政务云SRE用云优化能力中心制定总体规划；组建应急演练团队，华为侧一线客户经理及运维工程师协同政务云SRE团队成员组成应急演练团队负责与客户沟通和方案讨论。

• 演练准备

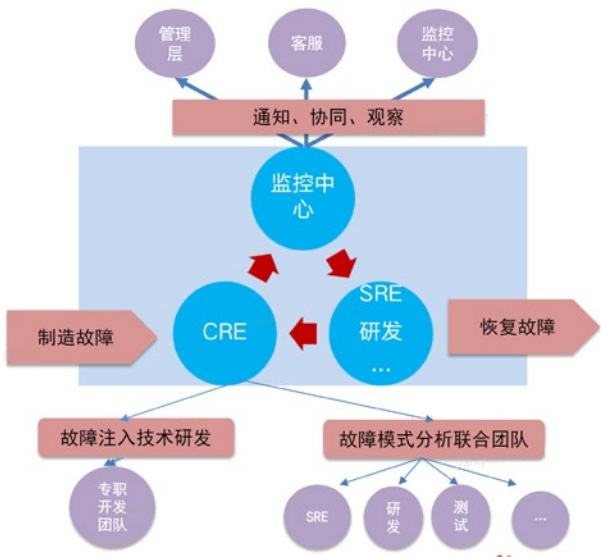
新上云时期或客户业务推广、版本更新、重大活动等业务场景，通过关键场景的切换演练，验证业务系统的稳定性，提前

识别关键业务性能风险点，提供业务调优建议，最终提升业务系统抵御故障的能力。

- » 确定演练计划：华为政务云SRE团队与客户方演练团队讨论本次演练设计范围，确定演练涉及场景，并确定最终演练时间与演练日期。
- » 确定演练方案&应急预案：政务云SRE团队确定各场景演练目标，并由用云优化中心输出演练方案初稿、监控告警方案及应急预案，通过华为侧演练团队内部评审后，提交客户方审核，通过后作为本次切换演练的演练方案与应急预案。

• 演练实施

- » 演练动员，政务云SRE团队在演练开始前，确认本次演练相关人员已做好演练准备，确保演练顺利进行。
- » 演练实施，由客户方启动演练，SRE工



程师配合按照演练方案实施演练步骤执行。

- » 观察记录，政务云SRE团队和客户记录每一演练步骤的时间与结论，并及时验证告警准确性。

• 演练评估和总结

对于整体演练过程及结果，客户和政务云SRE演练团队根据观察记录确定演练是否达到预期目标，最后由客户对结果进行审核。

SRE团队对演练过程和结果按照如下维度进行总结，输出演练总结报告：

- » 高可用方案/配置是否生效，若未生效是何原因，有何改进方案。
- » RPO和RTO是否符合要求，若不符合，是何原因及如何改进方案。
- » 对于应急预案是否适配或有无其他优化方向。

总结

数字化转型只有起点，没有终点；数字城市的安全和韧性，没有最强，只有更强。数字化时代序幕已拉开，政务云底座作为数字城市的核心支撑，稳定可靠、安全高效，是城市行“稳”致远谋发展的必然选择。十年磨一剑，政务云SRE团队不断

进行能力提升，通过一系列系统高可用架构优化以及应急演练能力优化方案和用云优化实践，实现业务安全稳定高质量，助推政务客户实现数字化转型。与此同时，政务云SRE团队还将800+政务云项目能力外溢，面向政务客户打造出咨询与规

划、上云与实施、运维与管理、优化与提升——“规、上、维、优、培”全场景专业服务产品和解决方案，为韧性安全城市建设持续注入动能，守护数字城市每分每秒，助力数字政府变革每时每刻，共筑数字中国美好未来。



租户业务架构韧性提升 保障上云用云 稳定可靠

作者: 晋彪

摘要

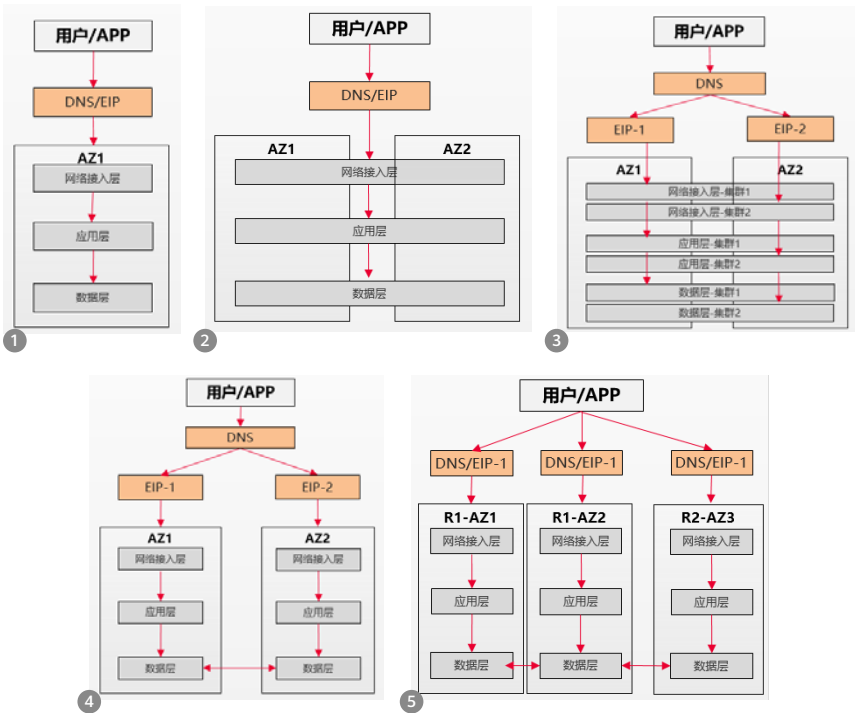
本文围绕大客户业务架构韧性展开，聚焦 CRE（CRE:Customer Reliability Engineering，客户稳定性工程师）团队通过高可用架构部署、弹性扩缩容、性能过载保护、故障可逃生、爆炸半径可控制等方式提升业务 SLA，以确定性方案解决不确定性风险，保障客户云上业务稳定可靠。

随着华为云的快速发展以及企业深度上云和用云，客户云上系统架构面临着系统可用性差、监控能力不足、故障恢复慢、业务运行风险未知等诸多挑战。面对基础设施和软件不可知、过载异常等不确定问题，华为云 CRE 团队探索出业务架构韧性的关键技术，通过高可用架构部署、弹性扩缩容、性能过载保护、端到端全链路快速逃生、爆炸半径控制等方式提升业务 SLA，切实保障客户上云，用云稳定可靠。

租户架构高可用

在不同的发展阶段，租户会根据成本以及环境属性对自身的业务架构进行不断迭代。衡量双活架构的3个关键指标为可应对的故障半径、系统恢复的时间、灾难发生时允许丢失的数据量。不同级别的可靠性架构，对成本的要求也是不同的，公有云客户需要基于业务发展现状，设计匹配自身业务的高可用架构模式。当前主流的租户架构形态演进态势为：①单AZ集群部署架构→②跨AZ高可用架构→③跨AZ双集群架构→④跨AZ双活架构→⑤两地三中心多活部署。

租户的高可用业务架构能够防范大部分硬件及软件故障，租户业务可用性可达到99.99%。在华为云，大部分客户使用②跨AZ高可用架构，部分大型互联网客户使用跨③跨AZ多集群架构/④AZ双活架构/⑤跨region双活等部署模式。



弹性伸缩

公有云租户均有严格的成本管理要求，对系统的冗余容量需要做精准控制，不允许大量闲置资源存在。如若发生大型AZ级故障，双活切换后极可能引入容量和性能问题，而极限弹性扩缩容却能弥补容量性能风险。

弹性伸缩根据客户的业务负载及策略，动态地增加或减少实例，从而实现云资源的最优使用，性能与成本的最优化。

弹性伸缩有两种实现方式：①自动伸缩，提前配置若干规则，当规则满足时，自动执行规则中定义的伸缩策略；②手动伸缩，手动在界面上改变实例个数。

弹性伸缩支持三种策略：告警、定时、周期。

» 告警策略：配置CPU/内存/磁盘/网络多个维度监控告警指标，随时根据业务负载有序伸缩。适用于业务负荷变化无规律场景。

» 定时伸缩：定时伸缩计划精确至秒级，适用于每天有明确的波峰波谷，或者有定时负荷突变场景。

» 周期伸缩：按天/按周/按月设定伸缩计划，适用于业务负荷波峰波谷存在较长周期的规律变化场景。

弹性伸缩的两种模式：①垂直伸缩，升级配置/降低配置；②水平伸缩：增加节点/减少节点。

弹性伸缩场景：EIP带宽弹性、ELB后端节

点伸缩、CCE工作负载伸缩、AS弹性伸缩ECS、高阶服务集群弹性伸缩节点数等。

弹性调度能力：管理，调度控制，规则编排；采集，监控指标收集，AOM/APM/CES；调度，容量/性能评估，弹性调度。

1 容器弹性扩缩容实践-容器弹性伸缩

【检测】检测条件：CPU/Mem、自定义指标、定时触发器、AI 触发器。

【计算】每个扩缩容条件设置阈值，突破持续周期；弹性扩容：取组合条件的或结果，如：CPU升至达80%或内存升至85%；弹性缩容：取组合条件的与结果，如：CPU降至50%，且内存降至50%。

【决策】控制扩缩容步长，即触发条件后，单次扩容的节点或pod数量。

【执行】扩缩容要启用优雅上下线，确保业务无感知。

2 故障场景下触发容器弹性扩缩容样例

某租户的容器化部署方案为双AZ双集群部署，即AZ1一套CCE集群，AZ2一套CCE集群，容器之上部署了一个调度系统。新建连接从网络接入层均分至2个AZ。假设AZ1发生批量故障，AZ1-CCE集群node-4/ node-5发生故障，流量均分至AZ1+AZ2的其他正常node节点的pod上。CCE检测到双CCE集群的pod负荷上升，触发扩容2个node，并且均分于2个集群，由于AZ1故障，AZ1的node扩容失败，AZ2的node扩容成功。调度系统检测到2个集群的pod负荷仍然较高，继续触发扩容2个node，仍然均分于2个集群，此时继续AZ1扩容失败，AZ2扩容成功。连续2个扩容步长后，业务流量3:7分布于两个AZ，流量由CCE的上层调度系统控制分发，最终达到AZ1发生故障，流量通过弹性伸缩的方式转移到了AZ2。

3 其他扩容能力分析

1) 使用弹性伸缩组 (AS) 实现集群节点的弹性伸缩

由于扩容过程中AS需要生成新的虚拟机实例，一般比较耗时（分钟级），适合流量趋势可预知的（如每天流量波峰波谷周期性伸缩）情况下的伸缩。

- » 动态告警伸缩：CPU/内存/磁盘/网络多个维度监控告警指标，随时根据业务负载有序伸缩。
- » 定时伸缩：精确至级定时伸缩秒计划。
- » 周期伸缩：按天/按周/按月设定伸缩计划。
- » CCE容器服务配置HPA，满足按需弹性。

2) 公网带宽伸缩：ELB可以绑定弹性IP（EIP），基于弹性IP实现带宽的弹性伸缩，华为云支持实时修改EIP的带宽规格。

3) ELB规格弹性伸缩：ELB自身弹性伸缩通过变更ELB规格来实现。

4) DCS服务支持下面几种方式伸缩：

副本伸缩：在主备部署、读写分离部署、Cluster部署模式下，均支持副本实例的伸缩。

分片伸缩：在Proxy集群、Cluster集群部署模式下，支持分片实例数的弹性伸缩。

5) DMS (Kafka) 服务支持下面几种方式伸缩：

- » Broker实例扩容：DMS支持Broker实例扩容，但不支持缩容。
- » Broker实例存储空间扩容：DMS服务支持Broker节点存储空间扩容，不支持缩容。
- » 服务规格变更：可更改服务规格，支持升配或降配。

6) RDS服务支持下面的伸缩方式：

- » 只读实例伸缩：RDS支持主从和读写分离，只读实例支持弹性伸缩。
- » 数据库实例规格变更：支持实例节点的规格变更，如CPU升降配、内存升降配、磁盘升降配。
- » 存储空间扩容：支持数据库存储空间自动扩容。



性能过载保护

1 实现目标

- » 可感知: 可以通过监控系统查看全系统的负载情况, 过载的时候有告警信息, 找到哪些压力源流量异常。
- » 压不死: 服务流量超过性能规格值后, 可以拒绝过载流量, 保证本服务在大流量下不崩溃。先要确保压不死, 把处理不了的流量堵在系统外。
- » 可干预: 可以通过租户维度、功能维度、资源池/集群维度快速识别异常流

量, 具体包括各维度的Top流量和突变流量, 并且支持对异常流进行手动或自动的干预。

- » 自愈恢复: 服务压力流量恢复后, 本服务功能可在快速自动恢复的问题。

2 关键技术

- » 过载限流: 过载限流 (ShedLoad) 是通过并对并发访问和请求进行限速, 或者对一个时间窗口内的请求进行限速来

保护系统的可用。一旦达到限制速率就可以拒绝服务请求, 一般通过客户端限流或服务端限流方式达成目的。

- » 过载降级: 在业务发生性能过载时, 对非核心业务做流控处理, 保障核心业务资源充足。
- » 过载熔断: 过载熔断即在服务性能过载时, 为避免雪崩效应, 采用中断服务的方式, 优先恢复服务状态。当依赖故障时自动熔断, 避免故障扩散, 更多采用过载熔断。

故障可逃生, 爆炸半径可控制

任何故障场景必须具备三板斧能力, 即故障后直接绕开定位、直接启动预案恢复、缩短恢复时长来提升业务SLA。租户业务全链路底线故障场景可逃生, 重点防范租户业务流上局部模块故障进而影响整体链路, 导致整体业务宕机。

链路上全部模块需要具备以下故障逃生能力:

- » 故障逃生能力: 通常使用假设法, 假设

链路上各个模块发生了不可快速修复的故障, 应该通过应急预案快速恢复, 而不是现场修复。

- » 业务冲击导致的爆炸半径: 从性能、隔离变化角度考虑将应用程序拆分。
- » 故障恢复需求: 更短的业务中断时间, 更低的人员技能要求, 更确定性的操作处理手段。
- » 典型恢复手段: 构建三板斧能力, 提升

应急预案价值, 使其成为快速恢复的基石。

- » 应急预案: 适配业务情况, 量身定制应急流程; 基于业务恢复目标, 建立联合应急团队来提升快速响应能力; 通过构建智能化应急恢复平台, 提升执行便捷性。
- » 故障预防保障措施: 提升可视化监控能力; 完善流量控制策略; 对应急预案演练有效性进行验证。

总结

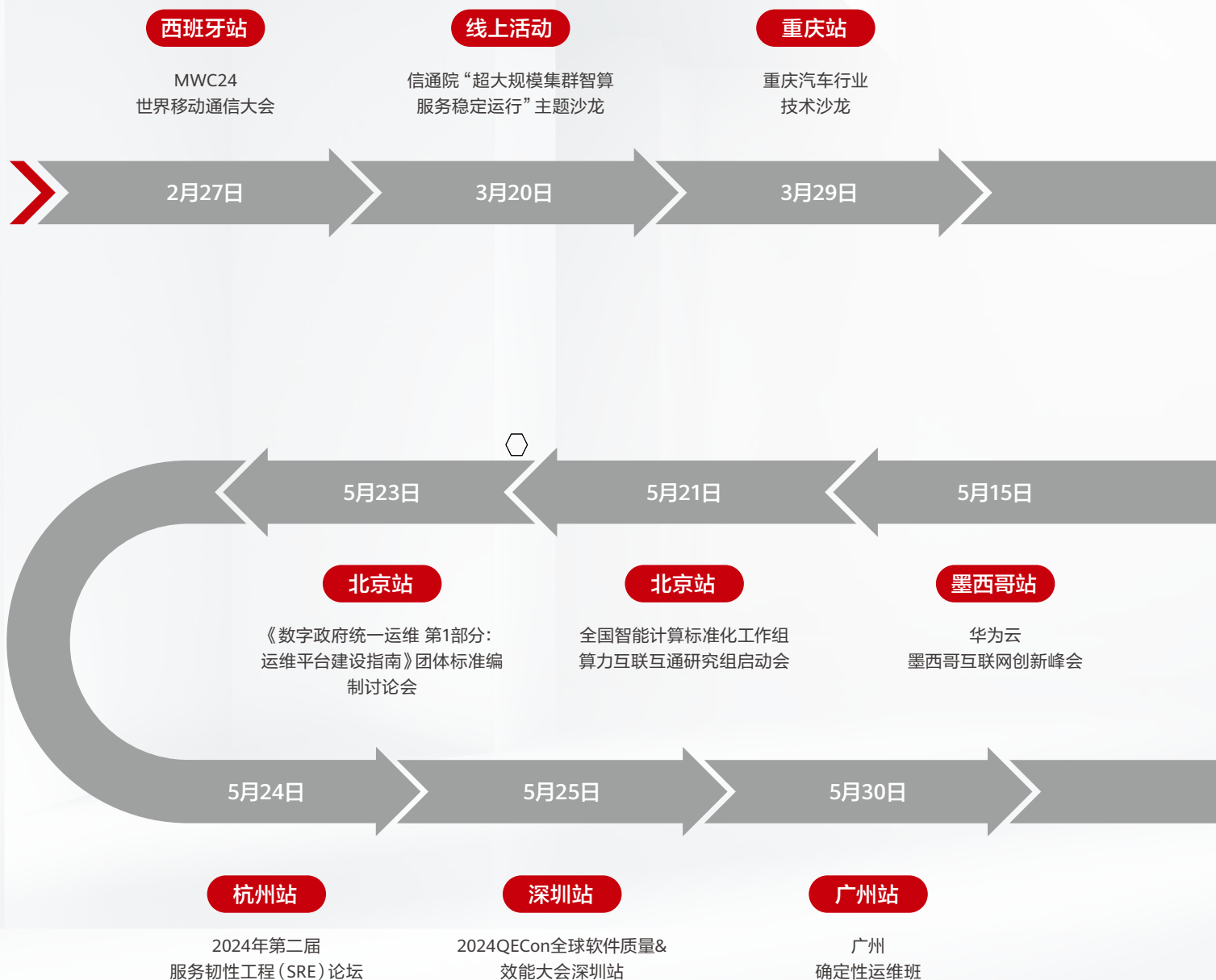
CRE团队通过最佳实践和最优方法论, 沉淀出客户业务架构韧性的关键技术, 从租户架构高可用设计出发, 适配用户不

同业务场景。以高可用架构为底座, 针对用户严格的成本管理需求, 对系统容量做精准控制, 为客户业务场景做性能保

护, 结合故障场景三板斧能力, 助力客户提升业务SLA, 以确定性方案解决不确定性风险, 保障客户上云、用云安全可靠。

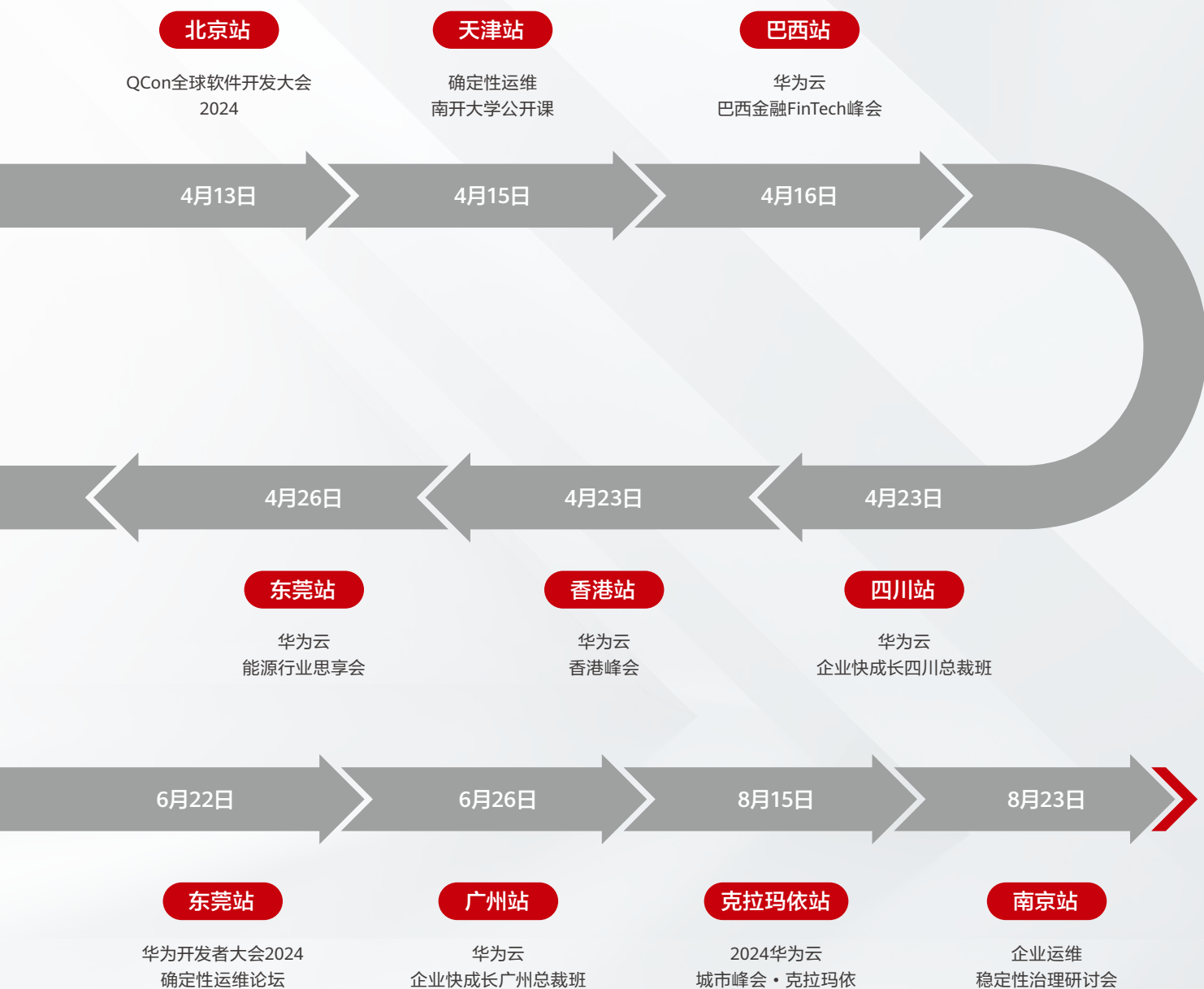


维享会精选活动回顾





2024年，维享会举行多次沙龙活动，汇聚行业大咖共同探讨AI技术的创新应用，围绕华为云确定性运维实践，进行思想碰撞和实践探索，通过确定性运维体系及解决方案，推动企业云上的业务创新，助力企业运维变革，帮助企业重构“管云”模式，护航千行万业行稳致远。



“

结合汽车行业的需求和痛点，围绕车联网业务高可用架构设计、可观测诊断、应急恢复能力和流量防护能力提升等，夯实确定性容量保障能力，分享如何帮助车联网企业构建确定性运维平台，完成从传统维护团队向IT平台服务的转型。

”

— 3月29日 重庆站 华为云SRE专家 谢文敏
现场分享车联网平台确定性运维促稳增效



重庆汽车行业技术沙龙现场

“

确定性运维公开课走进天津南开大学，围绕确定性运维体系及能力构建展开分享，与学生一起探索智能运维的前沿技术，培养新型SRE实用人才，与高校共同推进人才培养发展速度。

”

— 4月15日 天津站 南开大学确定性运维公开课
华为云SRE智能运维架构师 李开广
分享确定性运维体系及能力构建实践



确定性运维南开大学公开课合影

“

分享企业运维领域应用AIGC的实践案例，基于确定性运维的实践经验，提出以LLM为中心，结合多Agent协同的运维方案，并提出在大模型时代下，对下一代智能运维的思考。

”

— 4月13日 北京站 华为云SRE AI使能专家博士 张曦
分享确定性运维在LLM和Multi-agent在运维领域的
实验探索经验



QCon全球软件开发大会2024北京站现场

“

确定性运维全方位护航企业高质量发展，通过运维大模型加持，打造一站式专属云上运维管家。构建可观测能力方案，实现快速问题定位定界，使能千行万业管好云，帮助客户加速数字化转型。

”

— 4月23日 成都站 华为云专业服务专家 李伟
分享华为云安全稳定高质量实践经验



华为云企业快成长四川总裁班合影

“

基于华为云自身沉淀和大量实践形成的确定性运维体系，通过“确定性”的恢复及在线验证等能力，将运维团队从成本部门转化为生产力部门的实践，帮助业务管理者及其团队完成从“消防员”向“建筑师”转型，实现云上业务安全稳定高质量发展。”

— 4月26日 东莞站 华为云运维使能中心总监 林华鼎
分享华为云可用性保障“确定性运维”实践



“

本次论坛邀请上百名全球开发者到现场，会上围绕数字化转型带来的生产力转移展开讨论，明确需要确定性运维变革来助力数转，通过AI能力构建来助力变革。感谢来自学术界和企业的大咖们分享最新实践成果与方法，发布《确定性运维白皮书稳定可靠篇2.0》，与千行万业持续共创。”

— 6月22日 东莞站 华为开发者大会2024确定性运维论坛



白皮书发布合影



确定性运维专题论坛现场照片

“

华为云确定性运维安全合规体系框架保障业务稳定、数据安全、质量可靠、作业可信，全球化运维数据处理遵从本地法规，助力业务全球扩展，通过确定性运维实践经验和应用托管服务，助力出海企业应用可靠性提升，为中企业务全球扩展保驾护航。”

— 5月30日 广州站 华为云SRE运维高级专家 何元
分享中企出海运维合规及稳定性实践



广州确定性运维班合影

“

华为云与客户共同面向AI实践，基于AI原生应用引擎共创，使能确定性运维AI创新，通过深度整合AI技术与运维场景，为企业提供一种全新的、高效的运维方式，覆盖千行万业业务特性，加速实现智能化的运维决策和数字化转型。”

— 6月26日 广州站 华为云企业快成长广州总裁班



华为云企业快成长广州总裁班合影

“

从华为云保障云上克拉玛依安全稳定的经验与成果出发，依托华为云确定性运维能力体系与解决方案，基于“防范胜于救灾”的运维理念，通过“以防为先、防救相结合”的组合拳，保障数字克拉玛依安全、稳定运行。

”

— 8月15日克拉玛依站 华为政务云运维总监康镇分享华为确定性运维与安全保障克拉玛依城市安全稳定



华为云城市峰会克拉玛依站现场分享

“

结合客户在面临业务高速增长的智慧门店、业务出海等场景，分享基于确定性运维实践经验提升故障管理能力的方法。通过识别故障模式、构建故障快速恢复能力，提升系统韧性、在线验证等方式，帮助客户提高系统的恢复能力及故障管理能力。

”

— 8月23日南京站 华为云确定性运维故障管理专家王宏基分享故障管理的实践经验



企业运维稳定治理研讨会现场合影



精彩瞬间集锦



西班牙站
MWC24世界移动通信大会
华为云SRE专家黎果现场分享



巴西站
华为云巴西金融FinTech峰会
华为云拉美SRE主管王章玉现场分享



香港站
华为云香港峰会
华为云SRE高级专家董琳现场分享



北京站
全国智能计算标准化工作组
算力互联互通研究组启动会
华为云应用服务SRE总监郑磊现场分享



杭州站
2024年第二届服务韧性工程(SRE)论坛
华为云SRE运维高级架构师 刘涛
现场分享



深圳站
2024QCon全球软件质量&效能大会



企业运维稳定治理研讨会
华为云确定性运维故障管理专家 王宏基
现场分享



企业运维稳定治理研讨会
华为云AI原生应用引擎架构师 王永恒
现场分享

安全可靠 | 智能运营
资源高效 | 业务敏捷

华为云



确定性运维官网



维享会官网