智能运维，**云数据中心运维的未来之路**

◎ 马力/文

在云计算时代，IT系统建设越来越成为企业发展至关重要的一环。业务系统，以及支撑业务系统运行的基础设施通常是企业关注的首要目标；然而，保障业务健康运行的背后“功臣”——运维系统同样至关重要，因为每一次IT系统的转型，运维系统和业务保障都是最艰难的部分。在当前企业IT系统向云架构转型的时刻，运维系统再一次面临着新的挑战。

云架构对运维系统的新需求和新挑战

● 引入云计算和业务需求带来运维压力

随着越来越多的企业拥抱云计算，为了支持业务系统的快速上线、灵活伸缩以及更高的SLA要求，再加上有限的IT运维成本，运维人员将面临比以往更大的运维压力。在运维拥有海量设备且高度复杂的云数据中心环境时，如何提供99.95%或以上的高质量IT服务，提升效率并降低成本，是运维团队当前面临的最大挑战。

- **保障高运维质量：**云数据中心的设备规模从几十/几百向几万/几百万数量级演进时，海量硬件设备的使用对硬件故障的快速定位和隔离将带来巨大挑战；同时，采用虚拟化和分布式弹性技术也加剧了云数据中心的复杂度。这些都会导致运维难度增加，小概率故障成为常态且影响加大，用户级的99.95%或以上的服务质量承诺（SLA）很难保障。
- **提高运维效率：**虚拟化技术和众多开源技术的引入使得运维变得越来越复杂，传统人工运维模式处理速度慢、出错概率高。此外，传统人均50～100台设备的维护效率，在大规模云化环境下，需要投入大量人力。
- **保持低运营成本：**传统IT的资源使用率通常小于20%，在云化后资源使用率有所提升，但是个性化、按需弹性需求导致资源碎片化、负载不平衡以及扩容规划不精准，可能会造成整体资源利用率并没有达到规划目标，运维成本居高不下。

● 云架构用户体验保障和业务高可用带来运维的“不可知性”

为了提升资源的利用率，云架构下资源是共享的，而非独占，这与传统IT完全不同。云计算通过自动的弹性伸缩策略来实现资源共享与用户体验及业务可用性之间的平衡，这是云计算的核心优势之一。但这也带来了运维的新需求和新挑战，即运维人员往往并不知道业务

马力

实现 IT 系统全自动化运行的核心是建设一个智能化的运维系统，其核心能力包括 3 个方面：全生命周期自动化管理；智能化故障预防、发现与自愈；以及智能化容量运营。

系统具体运行在哪个硬件上，故障定位变得非常困难，解决这种不可知性要求运维系统要做到“更加全面的系统监控”，从而实现“可知性”。

● 传统IT系统和云架构IT系统的混合IT架构的统一运维管理

企业IT向云架构迁移不是一蹴而就的，而是一个长期共存的过程。两种架构导致运维工具差异大，对运维人员也带来了更大的挑战。如何实现两种IT架构统一、集中的维护管理，是运维系统面临的新课题。

● 全自动化要求运维人员的角色从“运维管理”转变成“运维研发”

分布式架构的云计算系统，其资源调度、业务伸缩、故障隔离和故障修复等都是自动化的，不可能基于人工来完成，这已经完全颠覆了传统IT的软件安装部署、业务使用和管理维护模式。因此，运维的工作不再是传统的运维管理，而是构建自动化运维模型和运维工具，这不但对运维人员、更对运维系统提出了新的要求。

智能化运维支持IT系统的自动化运行

实现IT系统全自动化运行的核心在于智能。系统具备完善的智能，才能够基于系统的状态、用户规模、业务体验质量和策略规则等，实现系统的弹性伸缩、故障隔离和故障修复等等，这一切都要靠一个智能的管理系统或者运维系统来完成。**系统的智能运维包括3个方面的核心能力：全生命周期自动化管理；智能化故障预防、发现与自愈；以及智能化容量运营。**

> 云数据中心的资源和业务规模都远远超过传统数据中心。传统手工方式实现云资源/云服务的生命周期管理时，效率低下、误操作风险高，自动化手段势在必行。>>

● 全生命周期自动化管理

云数据中心的资源规模和业务规模都远远超过传统数据中心。传统的手工方式实现云资源/云服务的上线、监控、升级、变更、扩容、限流、降级与下线的生命周期管理时，效率低下、人员误操作风险高，自动化手段势在必行。通过变人工处理为自动化处理，提升运维的人均维护效率，满足业务的敏捷要求，逐步向无人值守的自动化运维演进。

• **以工作流为中心的自动化作业平台，复杂操作简单化：**自动化作业平台提供了把日常运维经验标准化和工具化的框架，有利于运维经验的固化与共享。通过预先配置好使用频度较高的变更操作场景，比如已知典型故障的修复操作、资源池的扩/减容、补丁安装、健康检查、合规审计与不合规项整改、软件批量安装、管理节点的配置备份、配置信息提取，以及设备批量上下电等，可以实现开箱即用，将原本很复杂的操作简单化，从而大幅提升运维的效率，降低变更时人工误操作的概率。通过设置分权分域与提供操作日志，可以满足安全与审计的需求，实现可控、高效的运维变更操作。

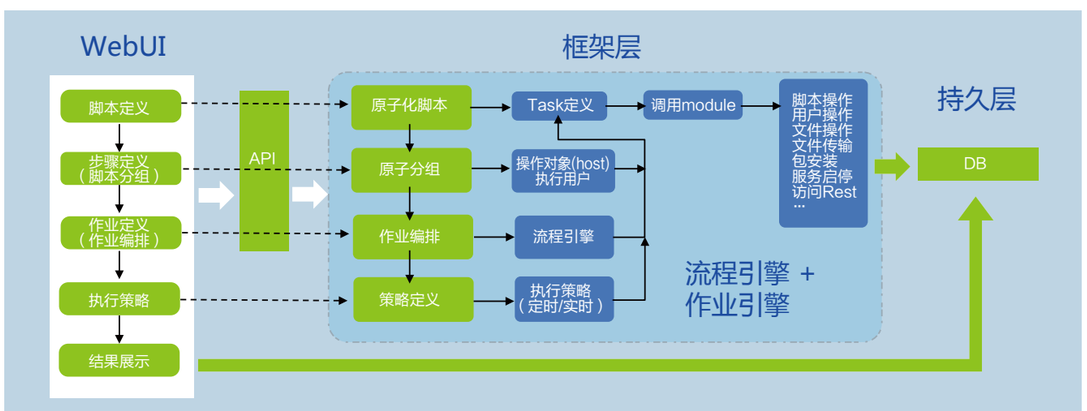

图1 作业平台业务流程

此外，利用平台提供的通用框架能力，运维人员还可以按需定制自动化作业。运维人员完成原子脚本开发后进行脚本可视化编排后提交，平台可以自动调度和分发执行，完成各种场景复杂作业的在线管理和自动执行。

• **标准化与一致性运维是基础：**由于传统数据中心里的软硬件“七国八制”，导致运维系统需要进行大量的兼容性配置，使整体建设的复杂度与难度倍增，难以落地。在云时代，通过使用标准化计算、存储和网络硬件，以及标准化软件的安装包、配置、权限、灰度发布策略、脚本和健康状态等，运维人员可以通过可视化、可预期的方式管理整个云环境，而且能够按照预设状态自行修正，解决传统数据中心内因为环境状态不一致所导致的频繁变更和人为失误等风险。

• **硬件即插即用，定期下线：**随着数据中心规模的增长，手工为主的硬件识别与安装方案将无法支撑资源的快速上线、扩容与下线。通过即插即用技术，只需要使用低技能人员将设备上架、上网和上电，运维系统就会根据该硬件的预期状态自动化完成端到端硬件系统的部署和上线；与此同时，通过云化隔离技术，硬件出现故障时也不再需要立即解决，只需让低技能人员定期替换即可。

• **软件一键发布，7×24永远在线：**随着敏捷、分布式软件开发部署模式的兴起，相对于传统数据中心，云数据中心内的系统升级变得更加频繁和复杂。通过一键式发布工具，实现从申请资源→发布部署→系统自检→自动化业务测试→回退/灰度上线的端到端自动化部署，同时支持全球多数据中心百/千级实例的集中发布。

• **移动运维：**手机端的运维App软件在手，专家可以随时随地移动运维，完成云资源的全生命周期管理。

● 智能化的故障预防、发现与自愈

传统模式下，运维人员的工作模式是被动等待问题发生，然后再进行故障处理。根据有关数据统计，运维人员平均每天计划内的工作只占50%左右，剩下的时间都是在到处救火。随着云数据中心规模快速增长，运维人员需要处理的事件量越来越大，人工救火将力不从心。这就需要一个智能的运维平台，利用大数据关联分析与机器学习技术为运维系统赋予人工智能，提供从故障预防到故障定位、再到故障闭环的智能保障能力。

> 随着云数据中心规模快速增长，需要一个智能的运维平台，利用大数据关联分析与机器学习技术为运维系统赋予人工智能，提供从故障预防到故障定位、再到故障闭环的智能保障能力。>>

• **主动故障预防：**故障处理再迅速也不如不产生故障，尤其是在大规模云数据中心场景下，即便很低的故障率也会产生一定规模的故障，为了避免到处救火，最好的方法是做好防火工作。

关键措施1：减少人工操作引入故障

根据华为公司IT部门的统计，变更操作是故障的导火索，超过50%的故障是由变更中的人工操作引发的。大多数的一级事故都由变更引起，主要原因是变更操作复杂，人工处理容易产生误操作。因此，通过变更自动化避免人工处理引发故障，是降低故障发生率的一个非常重要的举措。

关键措施2：系统亚健康智能分析，提前发现故障隐患

利用大数据技术，结合故障特征库进行跨数据领域关联分析，提前发现隐患、预测故障。与自动化策略执行系统集成联动，在用户发觉问题前将问题解决，避免对业务造成影响。

• **及时故障发现：**云数据中心由于技术堆栈层次多、技术架构复杂，如何识别故障是个很大的难点。构建一个从资源到租户体验端到端的监控体系，全面掌握系统运行状态

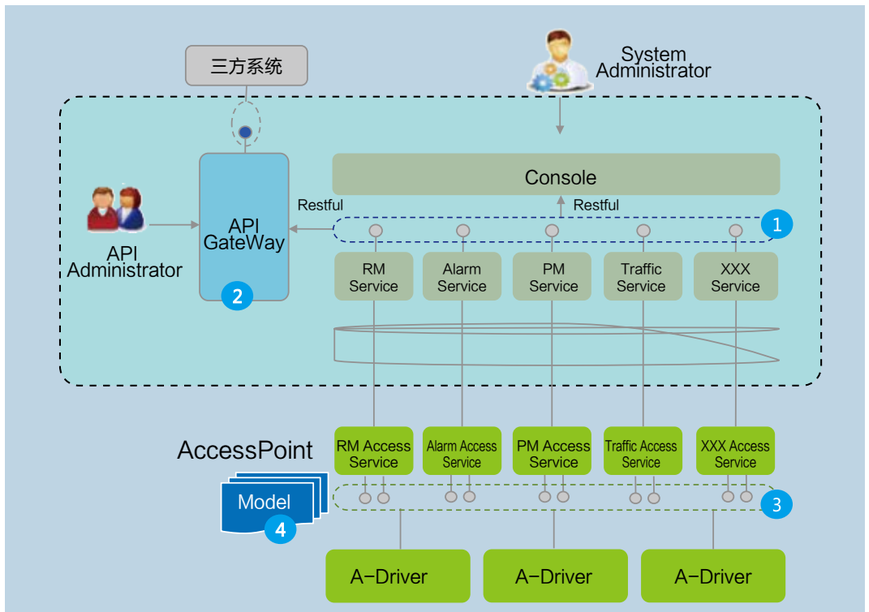

图2 开放的华为云运维平台

智能化的容量管理能够实现现状可视、问题可察、风险可辨、未来可测和调整可控，使云数据中心内资源的利用率提升到70%以上的水平。>>

数据，有助于准确识别出业务系统响应慢、查询速度慢、产品质量差（问题多、交易失败率高）和用户数量少/资源利用率低等问题的根源，推动技术团队不断改进，达到持续优化的运维管理目的。

关键措施1：构建全链路、主动、智能的全方位、多手段和多指标监控体系

运维系统需要支持从机房设施、物理基础设施、跨数据中心骨干网络、虚拟化资源池到云服务和应用的统一管理，实现多数据中心和多维度的集中监控。

当数据中心出现故障时，通过系统运行状态可视化，可以快速获取每个数据中心中资源和云服务的当前和历史运行状态，可以查看的信息包括性能容量、关联对象与告警，以及拓扑与各类日志信息。

关键措施2：系统运行状态可视化

在重点业务的服务运营保障中，通过可视化展示应用拓扑及其健康状态，可以使云基础架构与业务应用的各项运行指标和变化趋势一览无余。

通过提供各类运维对象的性能容量、告警统计与分析、资源利用率的报表，以及健康度和容量预测报告，IT运维人员与管理人员可以利用这些信息来支撑月度/季度的运维质量分析和年度IT架构规划。

- **智能故障定位：**云时代由于分布式和微服务化软件架构的流行，业务调用关系愈发复杂，出现故障后，对故障的快速定位形成了很大的挑战。

关键措施1：利用业务流跟踪系统快速故障定界

针对云服务微服务化后调用关系复杂和故障定位难的问题，需要有辅助定位工具来提高故障定位效率。通过对服务调用各环节SLA的监控来快速定位故障点，可以将故障定位的时间从小时级缩短到分钟级。

关键措施2：构建专家诊断系统，智能根因定位、已知故障自动化恢复处理

例行进行故障总结分析与持续积累，通过专家诊断系统将专家经验固化，可以实现故障定位的智能化和已知典型故障的自动恢复操作。

- **自动故障修复：**云数据中心规模的扩大带来了一个很大的问题——故障数量的提升。**根据华为自己的数据中心运维经验，一个较大规模的云数据中心，如果不进行故障的自动化归类和处理，每日各种级别的故障单可能超过上千个。**因此，迫切需要运维系统能够识别常见的故障，并有相关的故障自愈策略进行匹配。当故障发生时自动执行闭环策略，对于常见故障无需人工干预即可自动闭环解决。

	传统运维	OTT云化运维	运维效率提升
人均维护效率	50～100台/人	5000～10000台/人	100倍
资源使用率	小于20%	60～70%	3倍

- 智能化容量运营提升资源利用率

传统数据中心中，各业务部门独立部署的业务系统无法共享，服务器的利用率小于20%。数据中心云化后，云资源能够实现资源共享和动态调配，但同时也带来了碎片化、负载不均衡和SLA保障困难等挑战。

智能化的容量管理结合了大数据分析预测技术，将云数据中心内物理资源（如服务器、存储和网络等资源）和云资源（如虚拟机和块存储等）的实时容量视图、容量快照、负载现状和趋势，以及容量碎片呈现出来。针对资源负载不均的问题，传统运维平台因无法进行迁移/弹性伸缩而导致无法调整。而在云数据中心中，容量管理会向运维管理员提供低负载资源的分布信息，并提供缩减资源规格的建议；资源碎片化一般会导致20～30%“资源不可用”的情况，容量碎片管理向运维管理员提供各种资源规格的物理分布视图，并提供资源调整建议，提升现有资源的利用率。

云资源利用率达到一定阈值时，规划人员就需要考虑未来扩容问题。传统的容量预测主要依靠人的有限经验与数据来进行不可预知的扩容，往往会造成资源闲置率超过20～30%。而智能化的容量管理将资源的容量数据、应用行为分析、实际性能数据以及财务信息等相结合，对业务部门的关键应用对未来IT基础架构的各种资源容量的诉求进行高度准确和可靠的智能预测，向规划人员提供未来资源容量的趋势分析，供规划人员制定有效的采购和扩容计划，满足用户未来资源的高效利用。

智能化的容量管理能够实现现状可视、问题可察、风险可辨、未来可测和调整可控，使云数据中心内资源的利用率提升到70%以上的水平。

云数据中心运维的实践效果

运维比较成功的云数据中心，通过自动化和智能化的运维体系，面对百万级的服务器规模，在保障用户级99.95%甚至更高服务质量的前提下，实现了云数据中心运维效率的结构性提升：人均维护效率从传统人均50～100台提升至5000～10000台，效率提升100倍以上；而总体资源利用率从传统小于20%提升至60～70%，效率提升3倍以上（见上表）。

比如，**华为的研发采用云服务，通过标准化、自动化与智能化运维，目前已做到了11人维护10万台设备，资源使用率从10%以下提升至40～50%。**

同时，自动化、智能化和可视化运维平台的引入，使传统运维人员摆脱了以往机械式、重复性和低价值的日常工作，也最大限度地避免了人为错误的发生，间接保障了IT服务的质量，降低了运营成本。更重要的是，运维人员可以更多地投入到有价值和创造性的工作中，比如架构设计、开发以及新技术的评估和引入，以更好地支持企业的业务创新，更好地体现IT团队及个人在企业中的价值。

另外，通过自动化和智能化运维平台的引入，能够更好地通过工具的方式固化规范的IT运维管理流程。通过自动化流程的方式实现整体IT运维的规范性、标准化和合规性，以此保障对业务系统所承诺的服务质量（SLA），支持企业业务的健康发展。

华为云数据中心运维解决方案最佳实践

华为云数据中心运维解决方案除了帮助企业构筑一个自动化、智能化和可视化的运维平台外，还引入了华为多年来的实践经验，以及在新技术上探索的成果。

华为云数据中心运维解决方案除了帮助企业构筑一个自动化、智能化和可视化的运维平台外，还引入了华为多年来的实践经验，以及在新技术上探索的成果。>>

华为在运营商领域持续耕耘28年，已在全球建立了完善的技术支持体系，培养出了一批又一批技术过硬的专家，在IT领域可以复用这套全球化的技术支持体系。>>

● 运维经验沉淀、运维能力产品化

华为内部的运维团队负责维护着海量规模的华为企业云与私有云，月度进行运维质量分析、运维故障统计分析与经验总结，对于高危、重复度高的运维操作要求实现操作自动化。华为自营的企业云采用DevOps模式来快速构建和完善运维能力，经过充分验证后将运维能力进行产品化，纳入到华为云运维解决方案基线版本，保证华为内部运维的最佳实践可以批量提供给客户使用。比如前面提到的ECS服务调用链跟踪工具，就是日常运维经验沉淀的范例之一，通过整合到运维平台来不断提升运维能力。

● 能力开放构建云运维生态

华为提供了云运维的开发者社区，通过对外开放多层次API满足各类场景的应用开发需求，支持合作伙伴在云运维平台上持续积累、丰富运维的组件和工具，打造云运维的生态。

• 服务层的开放：所有服务Console使用的接口都对外开放，第三方可定制符合各行业场景的界面和Portal。

• 后台服务层的开放：所有运维服务通过统一的API GateWay对外开放，第三方可以基于接口开发新的运维工具，或对接第三方运维工具或系统。比如，基于开放的告警服务和资源管理服务开发本领域特有的业务拓扑视图，并实现业务节点状态的可视化；在混合IT架构下，性能容量、配置信息与日志都可以通过API GateWay对接客户自有的集中运维管理平台，实现全局共享一套运维体系。

• 设备接入层开放：提供南向驱动插件框架，第三方可以自行开发设备驱动，通过驱动管理服务动态接入新的设备对象，比如ZOHO开发的驱动已实现了非华为设备的监控上报管理。

● 微服务架构与容器化部署

华为云运维系统采用微服务架构支持容器化部署，具备良好的敏捷交付和可扩展能力。其中敏捷交付是指每个微服务都独立开发、发布和演进，可以快速迭代；易扩展是指每个微服务都可以独立部署并弹性扩展，保证了整个运维系统具备很强的扩展性，在小规模时可最小化部署，然后随着规模增长按需增加资源；而支持容器化部署，则大大削减了管理节点的成本开销。

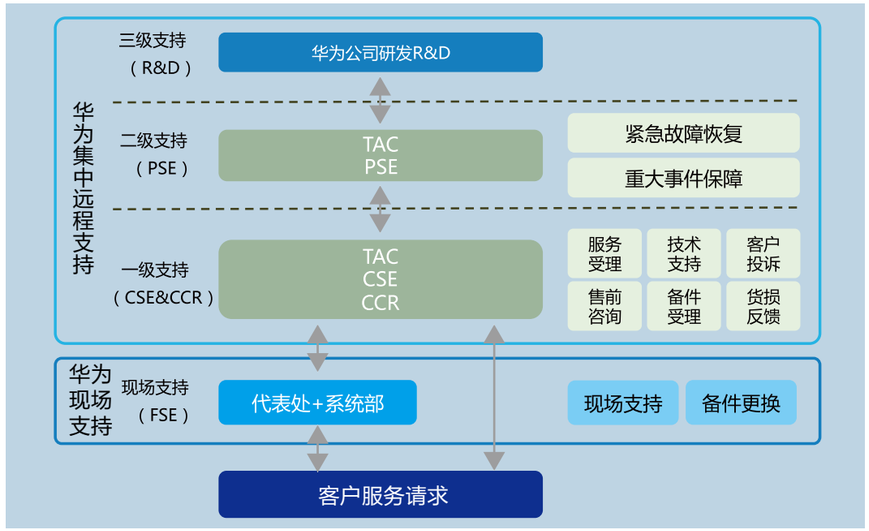

图3 客户服务中心业务服务体系

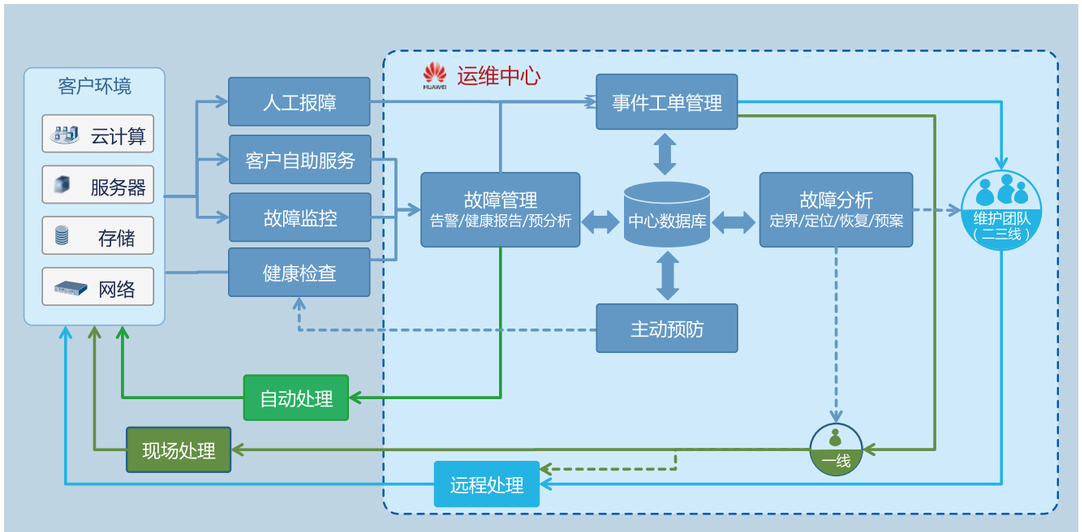

图4 IT运维体系全景

● 全球化的技术支持体系

华为在运营商领域（CT）持续耕耘了28年，已在全球建立了完善的技术支持体系，全球设有2个GTAC和多个RTAC，培养出了一批又一批技术过硬的专家，在IT领域可以复用这套全球化的技术支持体系。

华为提供多种运维模式供客户选择，包括客户自运维、华为现场代维或远程代维。客户自运维过程中遇到故障时可拨打7×24小时客服热线，同时也可选择部署CloudService实现自动报障，以及eCare全流程监控确保客户问题得到及时和有效的解决。

● 支持全栈式管理

借助在ICT基础设施运维领域的深厚积累，并充分利用自身产品线齐全的优势，华为提供了涵盖服务器、存储、网络、虚拟资源池、云服务和应用在内的完整的云数据中心管理能力，全栈的管理范围为端到端的业务监控、端到端的故障诊断定位，以及端到端的全生命周期自动化等能力的构建打下了基础。

华为将加大人工智能在云运维的投入与实践，让数据中心机器人融入更多的运维业务场景，替代传统的手工操作，提供高度自动化和智能化的“无人值守”式云数据中心运维解决方案。>>

近3年来，华为云数据中心的规模实现了数倍增长，但依托这套运维解决方案，在运维人员增长不到10%的情况下，SLA却达到了99.6%的水平，计算资源的平均利用率也达到50%以上，很好地支撑了研发业务的敏捷高速发展。比如，在2016年国庆假期的数据中心停电检修与版本升级变更中，涉及了分布在全国各地的11个机房、1.5万台物理服务器和30万个虚拟机，如果按照传统的运维能力计算，每位运维人员只能处理3000～4000个虚拟机，此次变更共需要投入100人才能实施完成；而借助智能化运维平台所具有的一键式上下电和批量版本升级操作能力，实际投入不到20人就完成了实施，每个机房上下电时长缩短了一倍（由10小时缩减至5个小时）。

云运维作为云计算必不可少的组成部分，会越来越展示出其重要性，成为云计算的核心竞争力之一。下一步华为将加大人工智能在云运维的投入与实践，让数据中心机器人融入更多的运维业务场景，替代传统的手工操作，提供高度自动化和智能化的“无人值守”式云数据中心运维解决方案。■