阿里云

弹性计算 ECS实例 **稳定性** 白皮书

目录

引言

每一个计算机从业者都希望能有一个绝对稳定的基础设施。这样，大家就可以把自己的精力专注在业务创新上。

为了这个目的，并提升效率，传统的 IT 公司会将整个技术部门分为两大类：基础技术部门和业务技术部门。基础技术部门通过横向技术平台为整个公司业务稳定性负责，业务部门为公司的业务创新负责。但是这种技术分工的前提是公司 IT 部门有足够的规模、技术和资金能力。中小型公司，创业公司，在没有取得业务成功和拥有足够的资金、技术实力之前，往往只能默默地承受。这就是 IT 业务发展隐含的技术门槛。

效率和普惠，永远是计算机发展的核心推动力。云计算的核心思想就是将资源和能力以服务的形式提供给客户，使得即使只有一两个开发者的初级创业公司也能简单地得到企业级的稳定性。

我们希望通过本文，读者可以系统性地了解到弹性计算（ECS）的稳定性如何定义，阿里云 ECS 产品做了哪些工作来提升产品的稳定性，以及作为用户，应该怎样做才能充分利用 ECS 产品提供的稳定性能力来为 IT 系统服务。

ECS 实例稳定性概述

ECS 简介

云服务器 ECS (Elastic Compute Service) 是阿里云提供的性能卓越、稳定可靠、弹性扩展的 IaaS（Infrastructure as a Service) 级别云计算服务。云服务器 ECS 免去了用户采购 IT 硬件的前期准备，让用户像使用水、电、天然气等公共资源一样便捷、高效地使用服务器，实现计算资源的即开即用和弹性伸缩。阿里云 ECS 持续提供创新型服务器,解决多种业务需求,助力用户的业务发展。更多介绍可参见阿里云官网。

ECS 实例稳定性定义

本白皮书讨论的 ECS 实例稳定性是指 ECS 实例运行状态相关的稳定性，用来刻画 ECS 实例持续稳定提供算力服务的能力。实例宕机、夯机等都是实例稳定性异常的表现。

ECS 实例稳定性建设初心

阿里云 ECS 作为提供算力的核心 IaaS 服务，承载了百万级客户的 IT 系统，阿里云深知客户将程序和数据部署在 ECS 上是源于对 ECS 的高度信任。ECS 实例稳定性是产品的核心特性之一，就好比一辆汽车的质量，不管其性能有多优良、价格多实惠，如果频繁发生故障，则一定会非常影响用户体验，甚至威胁到司乘安全。同样如果 ECS 频繁发生故障，导致客户业务或体验受损，也势必会辜负客户对 ECS 来之不易的信任。因此，阿里云 ECS 把实例稳定性当作产品立足之本，投入巨大资源进行稳定性能力建设，最终为客户提供稳如磐石的稳定性体验。

ECS 实例稳定性度量

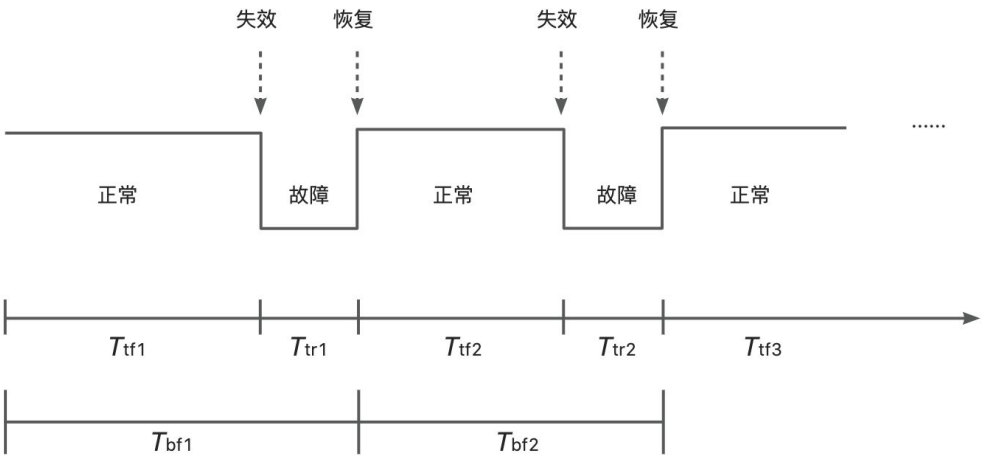
业界通用度量方式

稳定性是可信赖性的子项，分为可靠性、可用性和可维护性：

可靠性
在规定视角和条件下，系统内部件执行功能的能力。强调的是产品处于正常状态的能力。

可用性
系统投入使用时的可操作和可访问程度，或者实现特定功能的概率。强调的是产品处于正常状态占整个产品生命周期的比例。

可维护性
在规定视角和条件下，按规定程序和方法进行维修时，保持或恢复到规定状态的能力。强调的是产品恢复正常状态的能力。



稳定性时序图

以下是三个稳定性指标的量化算法：

指标	量化指标算法	指标解释
可靠性	$MTTF = \frac{(T_{tf1} + T_{tf2} + ... + T_{tfn})}{n}$	MTTF为平均失效前时间Mean Time To Failure
可用性	$A = \frac{MTTF}{MTBF}$	MTBF为平均故障间隔时间 Mean Time Between Failure $MTBF = \frac{(T_{bf1} + T_{bf2} + ... + T_{bfn})}{n}$
可维护性	$MTTR = \frac{(T_{tr1} + T_{tr2} + ... + T_{trn})}{n}$	MTTR为故障平均恢复时间Mean Time To Repair

阿里云采用业界通用的可用性指标度量 ECS 实例稳定性，这一指标符合客户使用体验的直觉。ECS 实例的可用性指标描述的是，实例处于正常状态提供计算服务的时长占实例生命周期的比例。基于可用性指标，阿里云 ECS 提供了明确的可用性 SLA（Service Level Agreement，服务等级协议），规定了服务质量和赔偿条款。

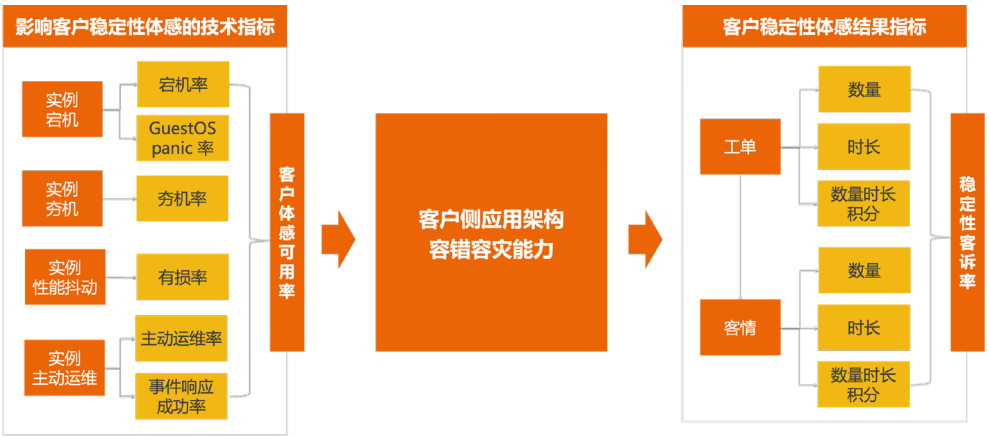
ECS 实例可用性，以一个自然月作为一个服务周期，计算方式：

维度	算法	SLA
单实例维度	$服务可用性 = \frac{（单实例服务周期总分钟数 - 单实例服务不可用分钟数）}{单实例服务周期总分钟数} \times 100\%$	99.975%
单地域多可用区维度	$服务可用性 = \frac{（单实例服务周期总分钟数 - 单实例单地域多可用区服务不可用分钟数）}{单实例服务周期总分钟数} \times 100\%$	99.995%

阿里云 ECS 于 2019 年 12 月 1 日率先发布了业界领先的 SLA-- 一个服务周期内，单台 ECS 实例的服务可用性不低于 99.975%、单地域多可用区服务可用性不低于 99.995%。直观来讲，该 SLA 约定了一个自然月内单实例不可用时长不超过约 10.8 分钟，单地域多可用区服务不可用时长不超过约 2.16 分钟。详见阿里云帮助中心云服务器 (ECS) 服务等级协议 (SLA)。

阿里云内部度量方式

以上是在与客户约定的 SLA 层面以“可用性”指标进行的度量，为了确保交付给客户的稳定性结果可预期，阿里云内部还会从以下更丰富和细致的视角去全面度量客户稳定性体感，核心指标如下图所示：



客户稳定性体感结果指标

通过“稳定性客诉率”指标来识别和跟踪客户反馈稳定性问题的情况，工单和客情分别代表了由低到高两种不同程度的稳定性客诉，工单和客情可以分别从数量、处理时长、数量 * 时长积分进行细化描述。

影响稳定性体感的技术指标

从技术视角量化可能影响用户稳定性体感的潜在因素,通过“客户体感可用率”指标来描述，其细分指标包括宕机率、GuestOS panic 率、夯机率、性能受损率、主动运维率和事件响应成功率等。通过细化技术指标的拆解与跟踪，可以清晰地诊断客户体感问题根因,从而进行优化改进。

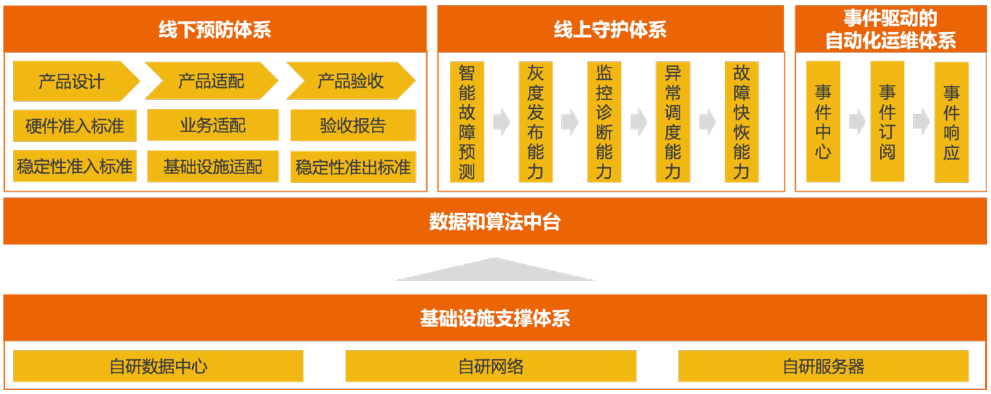
影响稳定性体感的过程

不同客户即使影响稳定性体感的技术指标一致，但也会因为客户的应用架构和容错容灾能力差异而产生不同的稳定性体感结果，这就是影响稳定性体感的过程。除了在阿里云 ECS 内部通过技术指标优化来保障客户体感，也会综合运用产品、服务等手段协助客户更好地使用 ECS。

ECS 实例稳定性核心能力

稳定性核心能力大图

ECS 实例稳定性能力体系整体上分成基础设施支撑体系、数据和算法中台、线下预防体系、线上守护体系、客户侧联动体系以及稳定性重保体系，如下图所示：



下面将对上述各项能力进行详细说明。

基础设施支撑体系

自研数据中心

阿里云自研数据中心为 ECS 提供了强健的基础设施能力：

环境动力

引入双向独立市电，机柜服务器 AB 双路供电，电池后备电源无缝接管，N+1 冗余柴油发电机自动切换接管，保证电力系统高可用。

冷却系统

阿里云数据中心通过技术创新，实现业界领先的浸没式液冷系统，液冷技术是指把 CPU、内存、芯片组、扩展卡等发热器件全部浸没于特殊的绝缘液体中，利用发热器件向液体传热，再由液体将热量通过单相循环或两相相变形式，最大效率地把热量传输出去的一种技术。阿里云通过长期的测试，验证了服务器浸入电子氟化液时的可靠性，通过大量多维度数据分析，显示基于电子氟化液的浸没液冷服务器各项指标均能满

足 SPEC 要求，各项电气特性和物理特性无显著变化，浸没式液冷服务器可以排除一些环境因素对服务器的影响，相比风冷服务器，设备失效率可降低 50%+。阿里云浸没液冷数据中心，获得了开放数据中心委员会（ODCC）和绿色网格（TGGC）的 5A 级认证，成为全国首座绿色等级达 5A 的液冷数据中心，液冷服务器技术还获得了中国通信标准化协会（CCSA）科学技术进步二等奖、工信部 2021 数据中心大会卓越创新先锋奖以及 2021 年 CCF 科技进步杰出奖等各类国家级奖项。

运维管理

阿里云数据中心建立了一套智能运维管理体系，实现了日常运维管理、现场安全管理、系统风险排查以及紧急环境应急等环节的体系化管理。监控和故障诊断方面，通过自研数据接入协议和天安监控设备，实现阿里云全球数据中心内设施的秒级数据采集，通过自研中心监控平台实现实时告警、根因定位和拓扑展示，结合中心运维管理平台和移动应用 APP，实现中心和本地 7*24 小时不间断 IDC 运维能力，机房管理员不论身处何地，都能提供分钟级故障分析处理能力。

自研网络

阿里云高可用网络架构以及自研核心交换机为 ECS 提供了稳定可靠的物理网络链路：

1. 高可用网络架构

阿里云物理网络实现了可用区 AZ 间低延时高速互联，AZ 内双冗余网络架构，数据中心超过三路由的线路保护，3+N 多线接入 BGP，实现业界领先的 BGP 弹性容量。阿里云服务器双上联边缘交换机，增强了服务器对边缘网络设备的容错能力。同时阿里云物理网络建设了集秒级监控、智能定位以及自动恢复于一体的故障自愈体系，可实现异常网络设备自动化隔离、异常流量自动调度以及异常路由自动优化等能力，大幅提升了物理网络的稳定性。

2. 自研核心交换机

阿里云实现了整数据中心核心交换机自研化能力，自研交换机运行的 AliNOS 系统具备高可用（软件特性简化、系统容错能力以及开放系统上的可运维性）、高性能（高性能管控接口、协议和芯片支持）、高敏捷度（软硬件解耦、网络架构高效迭代）的特性，同时 AliNOS 支持 AliCLI、RDMA、路由和基础系统性能优化、gRPC telemetry 和 Docker 热升级等能力。自研交换机的高可用系统架构、自动化测试以及高效的管控和运维支撑体系，实现了比商业设备更高的线上稳定性，上线以来连续多年平稳支撑了 618、99 大促和双十一等活动。

自研服务器

服务器作为数据中心基础设施的核心设备，承载着阿里云 ECS 的关键计算能力，也是 ECS 稳定运行的基石，2017 年以来，阿里云服务器全面采用自研方式，在稳定性方面具有核心优势：

1. 领先的自研工程化能力

服务器设计研发方面，阿里云通过深入自研底层硬件和软件，形成了自研的磐久服务器产品族（Alibaba

Cloud Server Series)，基于服务器硬件方升架构，以及倚天 710 CPU、CIPU、AliFlash SSD 和 Alimemory 内存等核心自研芯片及部件创新，从芯片部件级专研匹配业务场景，以实现更强更稳定的计算能力。

全链路质量保障方面，阿里云服务器通过严密的实验室及生产全链路硬件测试、智能化运营管理，实现专项测试前移，做到有效拦截硬件批次质量隐患，使得硬件非预期宕机率远低于业界水平。

2. 独一无二的超大规模海量场景打磨

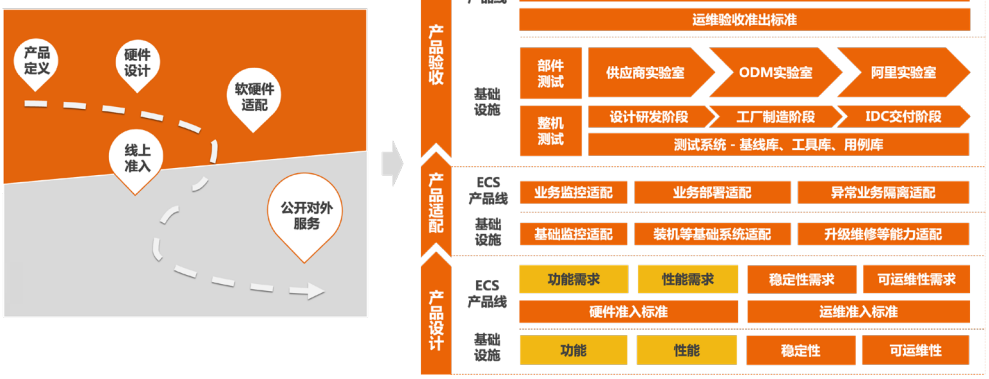
阿里集团核心 IT 系统构建于阿里云之上，阿里双十一超大规模极端大促场景，以及大促前地狱级全链路压测、无提前通知的混沌工程等对阿里云 ECS 提出了极大的考验，也是在此过程中，发现和修复了一般业务场景无法发现的问题，经过双 11 大促场景历练的 ECS 更加的稳定。

公共云百万级客户各类行业应用场景打磨，阿里云 ECS 承载了来自各行各业百万级外部客户的 IT 系统运行，包括：政府、电商、游戏、直播、视频渲染、AI 计算、HPC 等行业，这些行业客户的应用场景各不相同，为 ECS 提供了全方位的能力验收场景。

3. 和 Intel 等厂商的深度合作

经过阿里双 11 超大规模极端大促场景和公共云百万客户场景的打磨，ECS 稳定性沉淀了 10 年 + 百万级服务器的高质量打标数据，这些数据一方面作用于阿里内部系统，提升服务器设计研发能力，促进测试用例迭代，另一方面反馈给 Intel 等厂商，给厂商提供在实验室无法发现的故障场景，促进厂商产品质量迭代，形成良性互动的合作模式。通过此方式，我们和 Intel 等厂商建立了深度合作关系，CPU 微码升级等信息，Intel 等厂商会第一时间提供给阿里云，以便及时进行升级消除隐患风险，保障用户 ECS 实例稳定性。

线下预防体系



ECS 产品在上线公开对外售卖之前，会经历产品定义、硬件设计、软硬件适配、线上准入的完整线下生命周期。通过产品质量牵引，在这一阶段及早发现并解决问题进行风险预防，是提升产品稳定性的制胜关键，可以极大程度规避稳定性风险引入线上影响客户，并能够以较低的成本来进行产品稳定性改善。因而建立线下预防体系意义重大，其构成包括：

1. 产品设计

稳定性是 ECS 核心产品能力，稳定性需求和功能、性能需求一样，在产品设计的第一天就被提出，并且有明确的量化指标来衡量稳定性需求、可运维性需求、硬件准入标准和运维标准，稳定性需求涵盖 ECS 产品自身以及提供支撑的基础设施（服务器、IDC、物理网络等）。

2. 产品适配

基于产品设计阶段明确的各项需求，所有的线上异常需要能够自动发现、自动处理，力求对客户影响降低到最小，产品研发阶段不仅需要适配性能、功能，也需要适配监控、发布，考虑清楚每一个异常应该怎样隔离，怎样和业务适配，才能构成一个完整的对客户负责的产品。

3. 产品验收

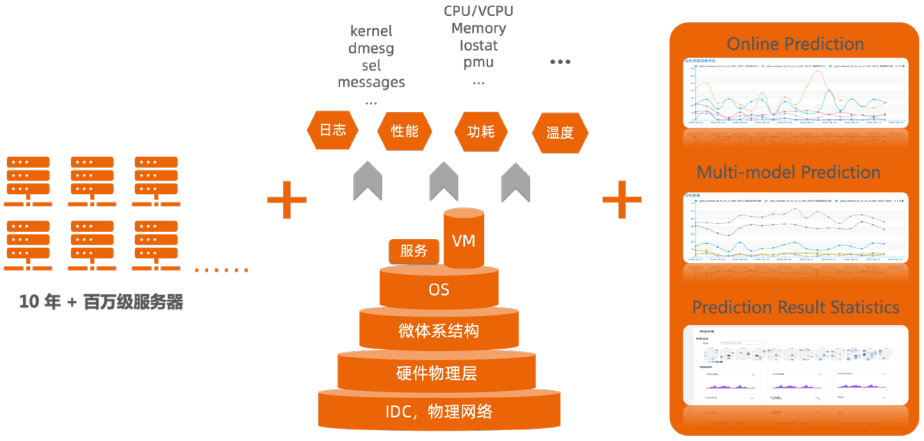
所有基础部件在进入 ECS 产品之前，都会经过充分的验证，在基础的功能和性能验证之外进行足够代表 ECS 产品特性的业务压测和可检测性验证。这些验证成熟后从产品团队前置到基础设施团队，实现测试和质量左移，产品团队和基础设施团队相互联动和促进。

线上守护体系

当 ECS 产品满足了准入条件，进入线上对外服务时，就需要提供持续的线上守护能力。其目标是最大程度降低非预期故障的产生，把可能的风险尽量转化为确定性的计划运维，而在非预期故障发生时能够将影响范围和时长控制在最小范围。围绕线上风险的预防、发现和处置构成了线上守护体系，其核心能力如下：

智能故障预测

ECS 宿主机硬件故障预测能力是进行主动运维保障实例稳定性的重要基础,为避险提供了时间上的提前量，配合热迁移或客户侧主动运维事件机制，可有效规避非预期故障。



海量高质量打标的生产数据　　精准全面的线上监控　　达摩院算法

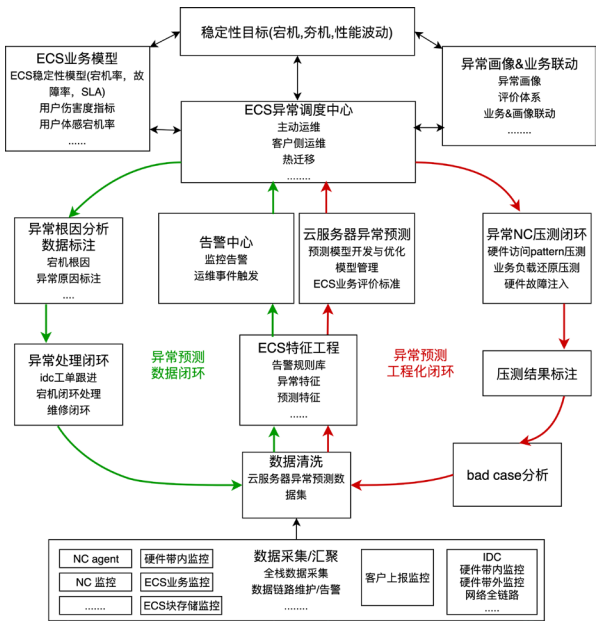
阿里云经过十多年发展，已具备百万量级服务器规模，沉淀了海量高质量打标的生产数据，通过对 IDC、物理网络、硬件服务器、微体系结构、操作系统、虚拟机和各组件服务的精准全面线上监控，并在达摩院算法加持下，形成了完善的故障预测体系，其特点包括：

高可靠性

多维度数据源综合校验，结合精准的预测算法，避免单一维度信息噪声，确保预测结果高可靠性。

高时效性

实时线上监控，通过高效的日志回传和大数据计算分析能力，确保预测结果的高时效性。



ECS 异常预测的核心环节包括：

全栈异常数据采集

由于 ECS 产品技术栈链路长，任何子系统的异动都可能造成 ECS 实例的异常，因此需要有全部子系统的完整、准确的数据作为分析基础。通过全栈数据采集和汇聚形成了基础数据层。

异常数据清洗

原始采集数据存在噪音，不能直接用于异常预测，需要进行大量的清洗工作，除了常规的排除异常值以外，还会结合业务语义，主动在业务流程中设计状态埋点，建立业务逻辑层的清洗标准。例如，排除测试、演练、搬迁等机器本身无故障场景的数据干扰。

ECS 特征工程和告警中心

常规监控告警依托全栈数据，基于专家经验、机器学习等建立告警规则并上报至告警中心。ECS 特征工程需要持续迭代，以排除异常告警阈值不合理带来的异常数据集污染。

异常根因分析和数据标注

异常根因分析和数据标注是异常预测的核心步骤，ECS 根因定位分析系统可以对 80% 的宕机根因自动分析，再经过专家人工确认来提升根因标注覆盖度。

异常预测工程化闭环

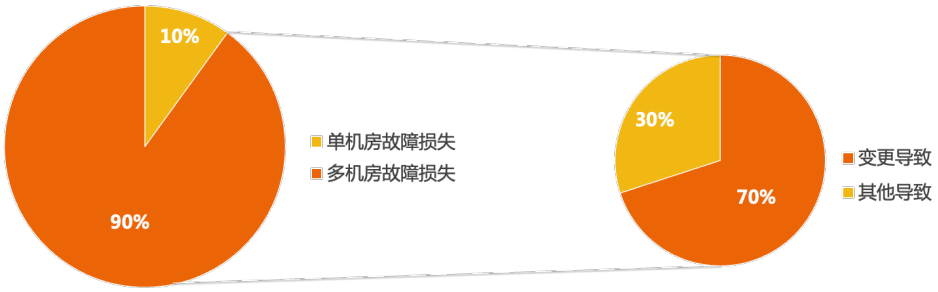
ECS 异常预测要建立一个能够持续进化的模型，首先要确保工程化的闭环是能够不断循环迭代的，其次要引入日常分析情况和评价机制，以建立对模型的及时反馈。异常预测模型数据结果交付给异常调度系统进行运维，还不构成完整闭环，因为宕机预测的结果可能存在一定的准确率、召回率，如果预测不准确的数据进入下一轮迭代则会严重影响模型正确性。因此必须有对预测结果建立是否宕机的事实标准，严格标记异常预测结果，并对漏报、误报持续构建异常情况库，有针对性地做特征工程，以便正确的数据输入到模型的下一轮迭代中。

ECS 业务模型生产管理

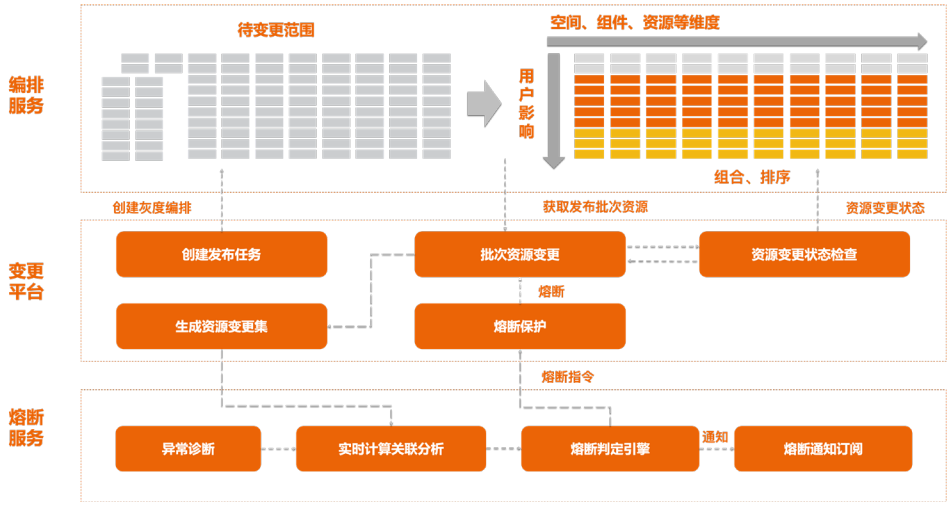
ECS 异常预测模型性能验证周期较长，往往需要好几个月。为了提升效率，在业务模型生产管理上引入了多模型、多版本管理模式，通过大样本进行并行验证，依据 ECS 业务评价标准自动进行评价。只有经过标准测试集验证和线上灰度，达到了性能标准的模型才能够最终应用于线上。且考虑到模型存在回滚的可能，旧模型的下线也不是彻底下线，而是在一个周期内逐步下线，在此过程中会保障预测数据的持续输出。

灰度发布能力

从相关统计数据看，多机房故障造成的损失相比单机房故障几乎高出一个数量级，而将近 7 成的多机房故障由变更引发，变更是业界普遍公认的线上稳定性“第一杀手”。阿里云 ECS 作为一个超大规模分布式产品，涉及的组件非常多，变更是常态化事情，做好变更风险防控至关重要。



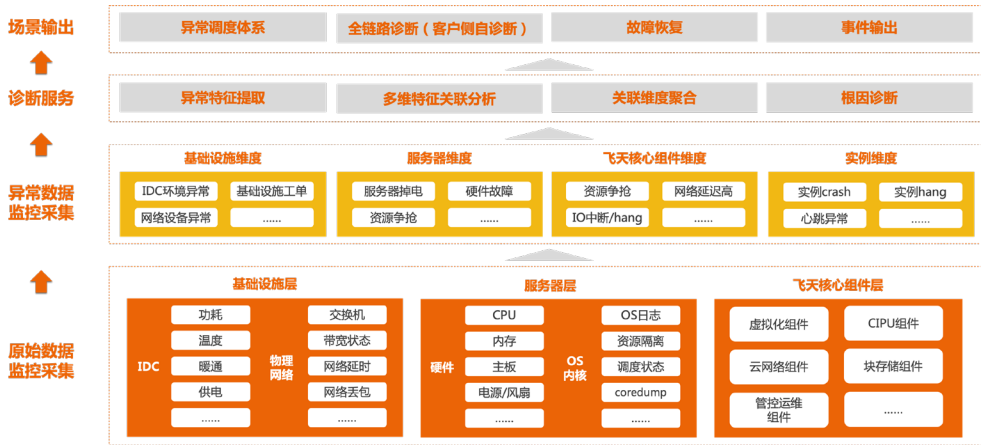
ECS 在变更风险防控方面，采用精细化的灰度编排和精准高效的异常熔断，结合变更平台对变更业务的流程管理，有效控制变更引入的异常“爆炸半径”，避免局部故障扩散给客户造成严重影响。



编排服务会对需要变更的服务进行分组和顺序的规划，按照空间范围、组件和资源等维度进行服务器分组，并设计合理的步长，即每一个变更批次包含多少台服务器。

通过变更平台来串联变更的执行和熔断控制。组件的一次变更首先在变更平台创建发布任务，并调用编排服务进行资源变更规划，生成需要变更的资源集合并执行具体变更动作。在变更执行过程中持续进行异常监控与诊断,当线上发生异常时会与执行变更的范围进行实时关联分析,当确定异常范围和变更范围吻合后，熔断判定引擎会产生熔断指令，变更平台响应该指令来暂停执行中的变更，从而实现变更异常风险拦截。

监控诊断能力



完善的监控与诊断体系就好比 ECS 系统的一双慧眼，能够全面地记录系统运行状态、识别系统异常，并且诊断根因，以便进行正确的异常处置。ECS 产品作为一个基础计算服务，依赖了 IDC、服务器、物理网络、虚拟网络、盘古存储等一系列资源，技术栈非常深，任何系统的异常都有可能表现为 ECS 实例的异常，而根因定位也随之变得非常复杂。

监控诊断体系由原始数据监控采集、异常数据监控采集和诊断服务构成，然后面向各种场景提供服务。

1. 原始数据监控采集

在 ECS 系统的基础设施、物理服务器和飞天核心组件层面进行最基础的原始运行数据采集，目标是提供海量全面的后端基础数据支撑，为更上层的诊断分析业务提供详实的依据。

2. 异常数据监控采集

在原始数据采集的基础上提取出大量细化分类的异常数据，将基础设施维度、服务器维度、飞天核心组件维度和实例维度过滤出异常数据，作为诊断服务的数据输入，从而降低数据量，使诊断服务可以聚焦在异常情况的关联分析。

3. 诊断服务

准确的故障识别和故障根因诊断，对于故障预防和故障止损来说至关重要，诊断服务会对异常监控数据进行异常特征提取，然后通过多维特征关联分析和关联维度聚合，推导出云服务器、物理机和控制面异常的根因，为自动化运维提供基础。

4. 场景输出

以上各层监控诊断能力支撑了各类应用场景：

事件输出

基于异常诊断服务生成事件，以事件形式解耦 ECS 内部运维系统之间的复杂关联，实现 ECS 系统内部自动运维；同时以 ECS 系统事件的形式，与客户进行必要互动，构建客户侧自助运维的能力。

实例主动运维

基于根因诊断形成的运维事件，可以进行宿主机下线、磁盘故障维修、ECS 实例宕机迁移或物理机强制重启等，最终实现对实例异常的自动化运维操作。

全链路诊断

对 ECS 后端全链路的日志信息进行数据打标与清洗，基于规则和模型建立起 ECS 异常表现和底层异常因素之间的关联，推导出 ECS 实例或物理机的故障根因，并在业务层面提供健康检查、批量诊断与告警通知能力。

客户侧自诊断

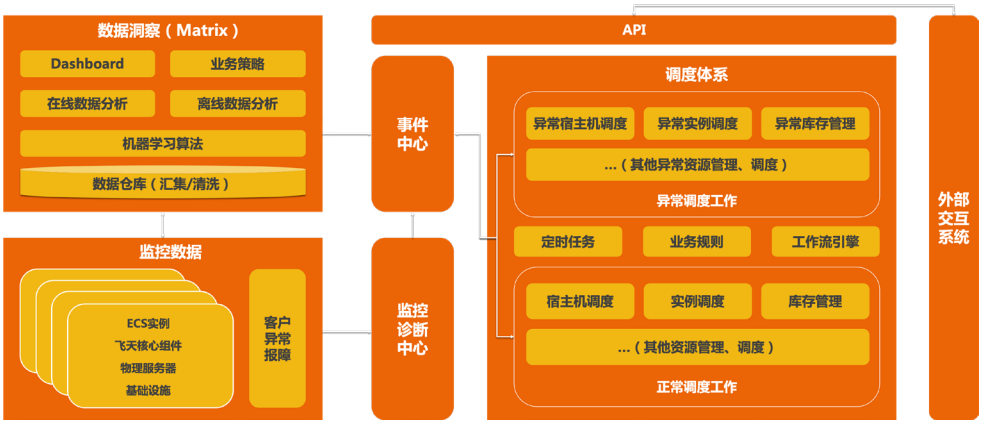
经过长期的能力沉淀，阿里云对 ECS 产品内部诊断能力进行封装，以客户侧自诊断能力产品化形式提供给用户，并集成在 ECS 控制台，帮助用户快速定位和修复问题，避免了创建工单寻求售后技术支持的过程，有效缩短 ECS 问题解决周期，提升用户体验。

故障恢复

在故障恢复场景中提供故障监控告警与批量实例的实时诊断，实现故障发现与根因定位，从而指导处置人员按照正确的故障预案进行处置。

异常调度能力

ECS 作为基础 IaaS 服务，其自身稳定性运维既有典型的复杂分布式系统运维的需求特点，又有云服务面向最终客户提供自助可定制运维策略的需求特点，一个外挂在 ECS 产品外围的传统 OPS 系统很难满足这些需求。ECS 从传统 OPS 体系演进为了“基于事件的异常调度体系”，同时满足了复杂系统的实时检测、与业务场景结合的实时运维决策以及将资源运维能力提供给客户的三大核心需求。



异常调度体系的核心如下:

1. 以事件为中心
所有系统间的交互以事件为驱动, 不管是客户侧告警还是基础监控告警, 都是资源的一个状态事件, 这些事件经过业务规则判断和加工后, 会形成内部运维事件或者客户侧事件, 分别交由 ECS 内部或客户处理; ECS 业务正常调度、异常调度以及与客户之间, 都通过不同种类的事件交互。系统之间实现松耦合, 可以快速扩展。

并且基于系统事件, 通过外部交互系统和 OpenAPI 与客户进行交互, 将 ECS 的运维能力提供给了客户, 其意义重大。

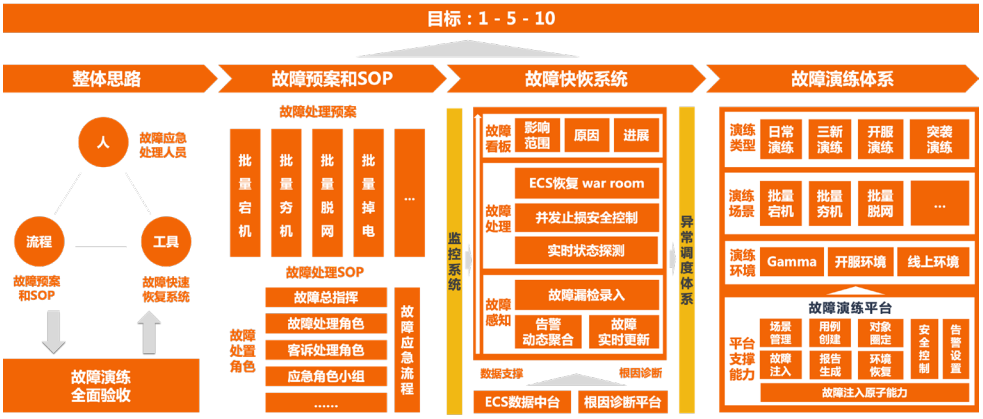
2. 统一的智能业务规则中心
在监控诊断能力的基础上, 松耦合利用 Matrix 系统 (数据中台) 的计算能力和机器学习能力, 根据异常现象与根因决策出合理的运维动作, 形成统一的运维业务规则, 并且能够对系统的运维业务数据进行挖掘而产生额外价值。

3. 整体运维体系和 ECS 产品调度体系融合
运维系统不再是外挂的外围系统,而是天生与 ECS 产品融合在一起,充分利用底层的大调度系统底座支撑, 运维体系整体逐步进化为 ECS 的异常调度体系。

故障快恢能力

即使阿里云 ECS 构建了完善的故障预测、故障隔离能力, 但故障仍不可 100% 避免。为了确保 ECS 可用性 SLA, 需要尽量降低故障恢复 MTTR (Mean Time To Repair, 故障平均恢复时长) 。阿里云内部对于故障应急处理有着 1-5-10 的目标, 即 “1 分钟发现 -5 分钟定位原因 -10 分钟止损”, 虽然结合不同故障场景处置耗时要求会有具体差异, 但这一目标给出了止损时长的方向性指导。

因此,ECS 构建了一整套故障快恢体系。其核心思路是对于人(故障处理者)、工具(快恢平台)、机制(故障预案)的持续优化, 并通过故障演练进行验收, 确保快恢能力持续迭代螺旋上升。



1. 故障快恢系统

快恢系统的目标是让故障处理人员能够安全、高效地处理故障，降低 MTTR。平台整体总体功能框架如下:

故障感知

当故障发生时，快恢系统第一时间感知到故障信息并进行通知报警。这里将会按照一定规则把同类型的异常实例和宿主机聚合，从而识别出是一次可能具有共性的批量故障，电话通知值班人员做应急处置。

故障处理

平台会实时展示故障实例、宿主机列表与根因诊断结果，并提供宿主机带外开机重启、启动实例、迁移实例等故障恢复原子操作能力。故障处置人员将会基于故障原因诊断，严格按照故障恢复 SOP，采用适当的操作方式进行故障处置。

故障看板

故障处理功能主要面向研发人员使用，而故障看板主要是提供给安全生产团队、客户服务团队等其他角色查看故障影响范围和恢复进展，以便在故障处置中协同进行客户侧服务对接。

2. 故障预案和 SOP

快恢系统是故障恢复的工具，而故障处理 SOP（Standard Operating Procedure，标准操作程序）则是故障恢复的流程规范，用来保障每次故障处理的质量。阿里云 ECS 从各种实际故障与演练中沉淀了多种典型场景的预案，诸如批量宿主机宕机、批量宿主机夯机、批量脱网等等。SOP 规定了参与故障处置的各方人员角色分工、故障诊断流程、故障恢复操作顺序，以及进展通报要求等。

3. 故障演练

阿里云 ECS 有完整的故障演练机制，以便验证快恢系统、故障恢复 SOP 的有效性，并充分锻炼故障处置人员的处置能力。会不定期发起有通知的计划性演练或无通知的突袭演练，在安全的演练环境中注入真实的宕机、夯机、脱网等故障，并严格按照 SOP 进行应急处置。演练完毕后会统计分析 MTTR，发掘 ECS 产品本身、快恢系统、故障恢复 SOP 以及人员操作方面的改进项，持续对故障恢复能力进行迭代优化。

事件驱动的自动化运维体系

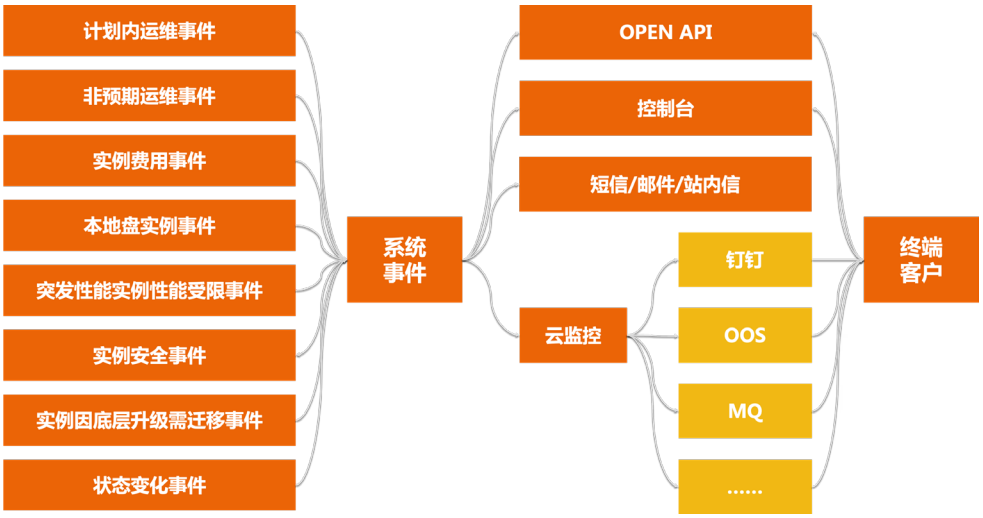
弹性计算 ECS 通过系统事件构建与客户的联动体系，客户可通过查询和响应系统事件与阿里云进行联动，具体描述如下:

系统事件概述

系统事件由阿里云定义，用于记录和通知云资源的信息，例如运维任务执行情况、资源是否出现异常、资源状态变化等。系统事件包括如下几种类型:

更详细的系统事件分类说明，请参考阿里云帮助文档 ECS 系统事件概述。

事件查询与响应



ECS 系统事件默认通过短信 / 邮件 / 站内信形式通知客户，客户可以在 ECS 控制台或通过阿里云 CLI 调用阿里云 OpenAPI 查询事件。也可以通过云监控进行个性化的触达订阅，包括配置电话、钉钉通知规则，或发布至 MQ 等供客户侧运维系统消费处理。查询、订阅事件后，可对事件进行响应，通过响应事件规避相关风险。

更多详细的查询和响应事件说明，请参考阿里云帮助文档 ECS 系统事件查询和响应。

ECS 实例稳定性最佳实践

以上介绍了通过产品技术和服务提供的 ECS 实例稳定性核心能力，而客户侧最佳实践将会指导客户用好 ECS，更加充分发挥 ECS 实例稳定性能力，保障用户业务稳定性。以下将从通用最佳实践原则和两个行业的具体最佳实践角度进行介绍。

按照用户使用 ECS 的全生命周期，通用最佳实践可划分为风险规避、问题容错、问题诊断几个大类。详细内容可参考下图：

以下将按照用户侧使用场景对上述最佳实践展开详细介绍。

风险规避场景

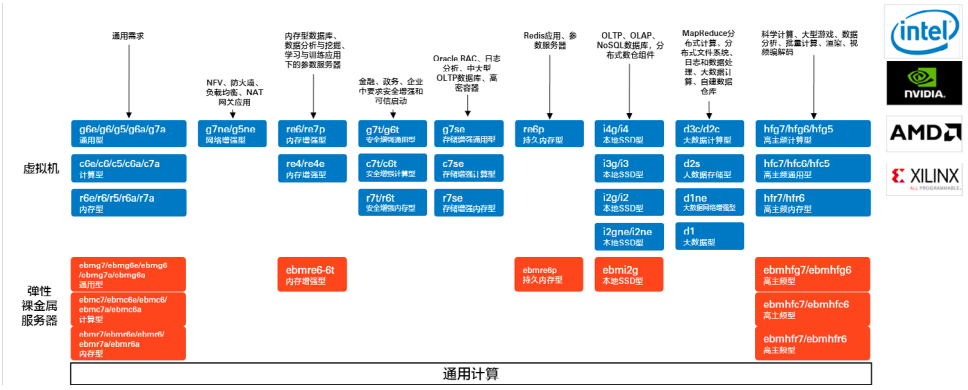
选择合适的实例规格

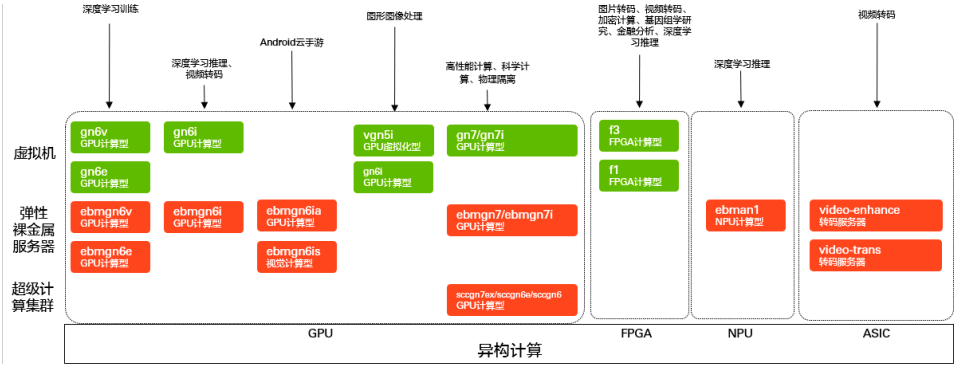
1. 业务场景

用户在购买一台 ECS 实例前，需要结合性能、价格、工作负载等因素，作出性价比与稳定性最优的决策。根据业务场景和 vCPU、内存、网络性能、存储吞吐等配置划分，阿里云云服务器 ECS 提供了多种实例规格族，一种实例规格族又包括多个实例规格。实例规格选型最佳实践，可帮助用户结合自身需求进行选型决策，并在库存不足、产品下线、使用抢占式实例等场景中，通过备选实例规格满足需求。

2. 方案描述

根据使用场景挑选：根据使用场景大致可以对实例划分为通用计算实例和异构计算实例，以下两张图分别列举了部分常见实例规格族和对应的业务场景，供用户参考：





根据典型应用挑选：如果用户使用的是类似于下图中的软件或应用，可以挑选右侧对应的实例规格族。

Web服务器	Apache	Nginx		计算型	c系列
中间件	SpringCloud	Dubbo	WebSphere	通用型	g系列
应用服务器	JBoss	Tomcat	Jetty	通用型	g系列
缓存	Redis	Memcache		内存型	r系列
数据库	MySQL	NoSQL		内存型/本地SSD型	r系列、i系列
大数据	HDFS	MapReduce	Spark	大数据型	d系列
AI机器学习	MXNet	TensorFlow	Caffe	GPU计算型	gn6v等

其他维度选型评估：常见的应用场景包括自建服务、通用场景、游戏服、视频直播、大数据、数据库、缓存、搜索、深度学习、图像处理等，详细的实例选型最佳实践参考帮助文档最佳选型实践 - 开始选型。

3. 涉及云产品列表

- 云服务器 ECS

弹性能力规划容量

1. 业务场景

用户的业务流量波动可能存在多种场景。例如，无明显的业务量波动场景，如果现有计算资源突然出现故障，会导致业务受到影响，很难及时进行故障修复或者替换；有规律业务量波动场景，每天固定时间业务量急速增长进入高峰期，到固定时间业务量下降，高峰期结束；无规律业务量波动场景，访问量突增和回落的具体时间难以预测，如热点事件对社交媒体业务的流量冲击等场景；以及复杂业务存在上述情况的混合场景。

2. 方案描述

上述场景在传统线下 IDC 模式难以应对，利用 ECS 的弹性扩缩容能力，结合实例按量付费，能够很好地解决上述场景问题。所需资源“用时可弹出，不用可退”，避免资源浪费，在确保业务系统容量安全的情况下，可以很好地优化成本。

弹性伸缩 ESS 自动管理容量：ESS 可以帮助用户自动完成弹性扩展与弹性收缩，用户可根据业务需求灵活配置自动化的伸缩策略。伸缩模式包括固定数量模式、健康模式、定时模式、动态模式和自定义等，其中动态模式支持通过 API 对接外部的监控系统。

固定数量模式配合健康模式可应对无明显业务量波动的场景，通过“最小实例数”属性，始终保持健康运行的 ECS 实例数量，自动替换不健康 ECS 实例，保证日常容量；定时伸缩模式，应对规律性业务量变化场景，配置周期性任务，定时增加或减少 ECS 实例；动态伸缩模式，应对无规律业务量波动场景，基于云监控性能指标（如 CPU、内存利用率），当性能指标高于设定阈值时自动增加 ECS 实例，反之自动减少 ECS 实例。

弹性交付效率优化：如果用户对于弹性扩展资源有较高时效性要求，可以采用 ECS 实例停机不收费模式，这是弹性计算产品的一个能力，停机后仅有存储产生费用，计算和网络不产生费用。在业务流量低峰期停机，释放实例的计算和网络资源，仅保留云盘存储资源，当需要快速扩容时直接启动 ECS 实例，即可节省掉最为耗时的存储资源创建过程。更极致的扩容时效性需求，可通过弹性容器实例 ECI 的急速启动能力来满足，最快的启动方式是采用 ECI 自定义容器组规格实例，秒级开通、秒级弹性伸缩。

详细的弹性伸缩能力使用方法参考帮助文档什么是弹性伸缩 ESS。

3. 涉及云产品列表

- 云服务器 ECS
- 弹性伸缩 ESS
- 弹性容器实例 ECI

响应 ECS 系统事件

1. 业务场景

当 ECS 资源出现异常、资源状态发生变化时，会产生系统事件通知用户。例如预测到后端物理服务器存在宕机风险,或者监控系统识别本地磁盘发生故障时,会通过短信、邮件、站内信等方式发送系统事件通知用户。此时用户需要与系统事件进行交互来完成运维。

2. 方案描述

ECS 系统事件是用来与用户交互的，用户收到系统事件后需要及时响应，以便消除实例风险。事件的查询和响应有两种方式，分别是控制台和 OpenAPI。

控制台方式

用户可以在 ECS 控制台查询待响应事件列表，并根据页面提示的操作进行响应。此方式存在人工运维成本，并一定程度上影响事件响应时效性。

OpenAPI 方式

推荐方式是通过 OpenAPI 来查询系统事件，也可以搭配云监控等方式主动订阅事件推送，然后由用户侧运维系统调用 OpenAPI 对系统事件进行自动化响应。此方式可降低人工运维成本，并确保事件响应的及时性，有效规避 ECS 实例稳定性风险。

查询和响应系统事件的具体操作方法详见官网文档。

3. 涉及云产品列表

- 云服务器 ECS
- 云监控

保持实例的操作系统更新

1. 业务场景

实例的操作系统是用户业务应用的运行环境，对于 ECS 实例运行态稳定性至关重要。如果长期不更新操作系统补丁，可能会遗漏重要的安全性或功能性升级，导致实例因操作系统 bug 产生 panic 宕机。而操作系统本身也存在维护生命周期的概念，如果使用官方明确不再维护的操作系统版本，可能会在遇到异常问题时得不到厂商或社区的官方技术支持，尤其是 Windows 之类闭源商业操作系统，问题的定位分析难度将会大大增加。

实例的操作系统属于用户管理范畴，用户需要管理好自身实例的操作系统环境的稳定性。

2. 方案描述

最佳实践方案是保持实例的操作系统的更新，包括：

保持操作系统版本更新

建议不要使用官方已经不再维护的（EOL， End of Life）操作系统版本。对于增量实例，在购买 ECS 实例时，请勿选择已明确标注 EOL 的镜像；对于存量业务实例,需要对 EOL 版本操作系统评估并更新。解决镜像 EOL 问题的方法详见官网文档。

保持系统补丁和驱动更新

系统补丁和驱动软件通常会包含重要的安全特性或功能特性升级，建议保持系统补丁和驱动更新，以确保系统稳定性和安全性。

针对各个操作系统的新特性、安全补丁等，阿里云会定期更新公共镜像的版本，详情请参见公共镜像发布记录。此外，阿里云会定时从开源社区官方或者操作系统原厂同步至阿里云镜像站中，用户可以按需更新新特性、安全补丁等。

3. 涉及云产品列表

- 云服务器 ECS

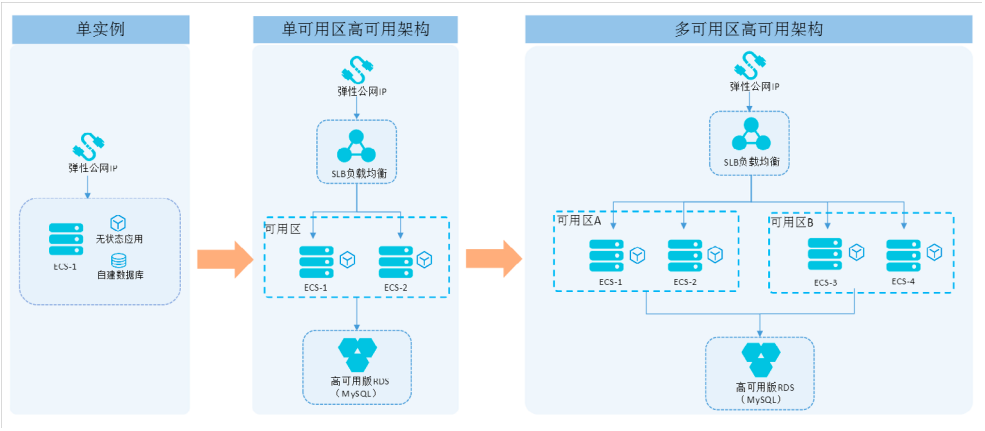
问题容错场景

建设高可用架构

1. 业务场景

若用户仅采用单个 ECS 实例部署应用，会出现 ECS 实例宕机后服务不可用的风险。再如，业务虽然采用多个 ECS 实例部署应用，但 ECS 实例全部部署在同一个可用区内，也可能出现可用区粒度的故障导致用户业务整体不可用。针对上述场景，需要构建 ECS 实例高可用架构。

2. 方案描述



业务仅采用单个 ECS 部署应用，不具备高可用能力，不推荐在生产环境内使用。

单可用区高可用架构

为实现单实例容灾，可在同一可用区内采用多个 ECS 实例部署相同业务应用，并将这些 ECS 作为 RS（real server）挂载至 SLB 负载均衡实例后端，负载均衡实例与弹性公网 IP 绑定来对外提供服务，形成上图所示单可用区高可用架构。ECS 部署的业务应用需要具备无状态的特性，当其中一个 ECS 实例宕机时，业务请求可以通过 SLB 实例分发至其他健康 ECS 实例，继续无差别地提供服务。

多可用区高可用架构

为实现单可用区粒度故障容灾，需要在上述架构基础上，将 ECS 分布至多个可用区，通过 SLB 完成可用区之间的负载均衡、流量切换。同一可用区内的业务模块需要完整独立部署，不存在跨可用区的互相调用，避免形成可用区之间“拧麻花”式耦合，以便发生可用区粒度故障时能够彻底将流量从一个可用区切换至容灾可用区。

3. 涉及云产品列表

- 云服务器 ECS
- 负载均衡 SLB
- 弹性公网 IP
- 专有网络 VPC

建立应用防抖能力

1. 业务场景

网络或 ECS 实例自身性能发生短暂恶化并恢复，通常被称为性能抖动，一般时长在秒级或分钟级左右。性能抖动可能造成用户业务请求超时失败，而造成抖动的可能因素非常多，包括 ECS 后端问题、运营商网络链路问题以及用户实例的操作系统内部问题等，现实情况下性能抖动难以 100% 规避，因此建议用户从自身应用层面建立防抖能力，保障业务请求成功率。

2. 方案描述

针对 ECS 部署的应用建立防抖能力，不依赖其他云产品能力，而是建议用户在应用层配置合理的请求超时、重试机制，一般涉及如下配置：

超时配置

如果业务希望首次请求的成功率达到某个水位（例如99.9%），即不进行请求重试的情况下满足成功率目标，可以对线上请求进行响应时长分位值统计，并将对应分位值耗时作为请求超时配置。

重试配置

通过增加请求重试配置，可以有效提升请求成功率，假设单次请求的平均成功率为 90%，增加一次重试，理论上请求成功率可以提升至 100%-(100%-90%)2 =99%。同时建议设置最大请求次数，避免过度重试造成资源浪费。

重试间隔

当遇到偶发短暂网络拥塞或断网等场景造成的请求响应时间升高，若发生请求超时后立即进行重试，可能异常因素还未消除，导致重试请求仍会失败。因此，建议配置一个合适的重试等待间隔,例如第一次失败后等待1秒，后续每次重试等待时间增加1秒，直至最大重试间隔设置，具体策略可结合实际应用需求配置。

3. 涉及云产品列表

- 云服务器 ECS

选择合适部署方式

1. 业务场景

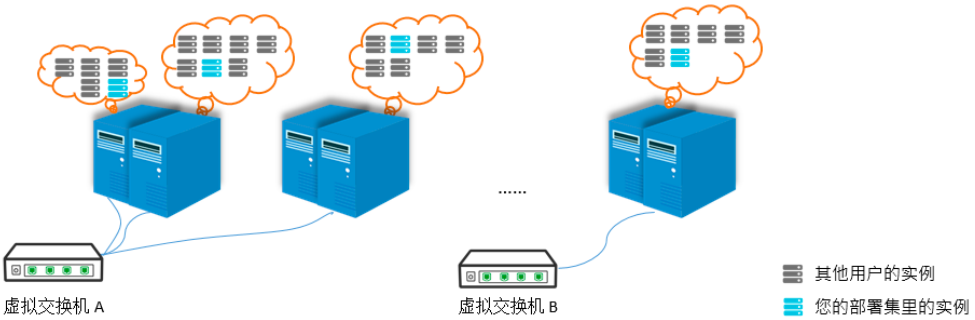
用户应用采用分布式集群化部署方式，同一组应用由多个 ECS 实例承载，应用具备单个 ECS 实例宕机容错的高可用能力，但是需要规避同一应用的多台 ECS 同时宕机，例如数据库主备服务器、分布式存储服务的相同数据分片副本等等。此时,需要对 ECS 所在物理服务器进行打散,避免物理机宕机影响多个 ECS 实例。

与一般业务相比，某些特殊场景（如游戏行业）对实例间互访有着更高的性能要求，希望将互访的实例部署在同一物理服务器来实现性能保障。

以上场景都可以通过合适的部署方式来满足需求。

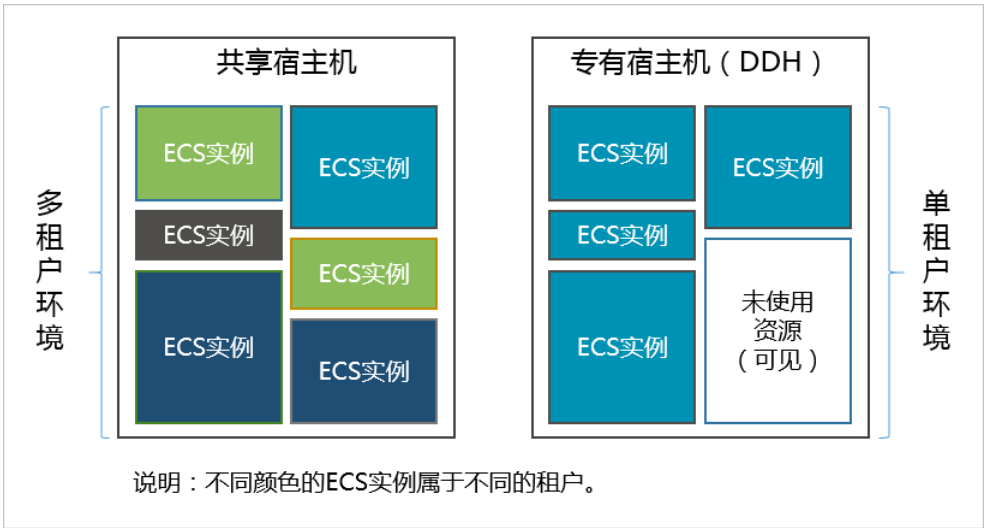
2. 方案描述

部署集实现打散：部署集是控制 ECS 实例分布的策略，使用户能在创建 ECS 实例的时候就设计容灾能力和可用性。使用部署集将业务涉及到的 ECS 实例分散部署在不同的物理服务器上，以此保证业务高可用性。在部署集内创建 ECS 实例时，会根据用户事先设置的部署策略，分散启动指定地域下的 ECS 实例。下图是利用部署集能力提升业务可靠性的典型示例，用户的四台 ECS 实例分布在四台不同的物理服务器上。



部署集的使用方式详见帮助文档部署集概述。

专有宿主机实现性能保障：专有宿主机 DDH（Dedicated Host）是指由一个租户独享物理资源的云主机。利用专有宿主机，用户可以获取 ECS 实例与物理服务器的拓扑关系，并进行灵活自主的资源部署规划，能够将需要互访的实例按需部署在同一台物理服务器，来满足实例间互访的更高性能要求。专有宿主机具体使用方式详见帮助文档什么是专有宿主机 DDH。



3. 涉及云产品列表

- 云服务器 ECS
- 专有宿主机 DDH

问题诊断场景

实例健康诊断

1. 业务场景

当用户在操作实例过程中遇到问题，需要进行针对性的问题诊断以寻求修复方法，例如实例无法启动、实例无法远程登录等。另外，在日常运维中，需要全面了解实例整体的健康情况，以便及时发现并处理异常情况，避免影响业务。

2. 方案描述

实例健康诊断功能可以对实例的系统状态、网络状态、磁盘状态等进行全方位的诊断，帮助用户了解实例的健康状态，及时发现并解决常见的问题。可覆盖的诊断场景包括实例性能问题、实例无法连接或启动异常、网络问题、实例操作未生效、资源配额不足、费用类问题、安全风险检测、实例费用及安全行为审计等。

实例健康诊断是免费的服务，可以通过控制台或 OpenAPI 发起诊断。详细操作方法参考帮助文档实例健康诊断（控制台）、实例健康诊断（OpenAPI）。

3. 涉及云产品列表

- 云服务器 ECS

网络连通性诊断

1. 业务场景

当用户的 ECS 实例、弹性网卡、公网 IP 等云上网络元素之间发生连通性异常时，需要定位网络不通的原因，以便排除故障。

2. 方案描述

网络连通性诊断可对云上元素之间的网络连通性进行定位分析。用户可通过 ECS 管理控制台 - 自助问题排查 - **网络连通性诊断**页签发起，通过指定诊断线路、发起诊断任务、查看诊断结果 3 个步骤即可完成自助诊断。指定诊断线路时，明确了 VPC、网络通讯的发起端和目的端、目标端口与协议，即可对通信链路进行端到端的连通性诊断。详细操作方法参考帮助文档网络连通性诊断。

3. 涉及云产品列表

- 云服务器 ECS

互联网行业最佳实践

Web/APP 应用是互联网行业的典型应用场景，常见于电商、社交、文娱等行业。业务负载方面具有一定周期性和间歇性高并发的特点，例如日常业务在自然日内负载呈现峰谷效应，而在日、周、月度表现为相对稳定的同环比特征，在电商大促、社交媒体热点事件等特定场景下又呈现为突发业务负载峰值。因此，在架构层面通常采用分布式架构以便支撑高并发负载，而成本控制方面又有资源弹性管理需求。

虽然互联网行业通常具备业务层面的容错能力，但 ECS 的非预期宕机仍然可能带来业务的抖动，特别是业务高峰时间段，一旦出现 ECS 宕机可能导致请求到该节点的用户受到影响。如电商秒杀场景可能下单失败、社交场景可能消息延迟等等。这类异常非常影响用户体验。阿里云经过众多互联网项目的沉淀打磨，总结了一套 ECS 在互联网行业应用场景的稳定性最佳实践。

以下结合互联网行业特点，对通用最佳实践中典型场景的落地展开详细介绍，并进行一定扩展。

选择合适的实例规格

互联网应用常常会涉及多种业务或组件类型，需要结合具体业务特点进行合适的 ECS 规格选型，常见业务的 ECS 规格选型建议如下:

业务类型	应用举例	推荐选型
Web服务器	Apache、Nginx	计算型，如：c系列
应用服务器	JBoss、Tomcat、Jetty	通用型，如：g系列
中间件	SpringCloud、Dubbo、WebSphere	通用型，如：g系列
缓存	Redis、Memcache	内存型，如：r系列
数据库	MySQL、NoSQL	内存型，如：r系列
大数据	HDFS、MapReduce、Spark	大数据型，如：d系列
AI机器学习	MXNet、TensorFlow、Caffe	GPU计算型，如：gn6v

弹性能力规划容量

基于分布式架构和合适的规格选型，结合业务负载特征分析，用户可以在云上进行方便的容量管理。以核心后端在线计算服务为例，常态下应对稳态负载可以采用非共享型实例，配合固定模式、健康模式弹性伸

缩策略可采用动态模式弹性伸缩策略,弹性扩展突发性能型实例来应对业务量高峰或无规律的突发业务量。弹性资源可以做到快速伸缩,“用时可弹出、不用可退”,简化容量管理、兼顾 ECS 稳定性与资源成本。

响应 ECS 系统事件

对于任何使用 ECS 的业务系统,都强烈建议用户及时响应 ECS 系统事件,自主决策计划性运维时间,避免实例风险恶化。推荐使用 OpenAPI 查询和响应事件,例如通过云监控配置需要订阅的事件类型及通知方式,包括邮件、短信、钉钉通知和订阅至业务侧消息系统,在用户侧运维系统响应事件之前可通过 SLB 将相关 ECS 实例的流量摘除,然后通过 OpenAPI 响应事件,待事件执行完毕相应运维动作后(如实例 reboot),再恢复相关实例流量,这样可以自动化且优雅地完成实例运维,有效降低事件执行期间的业务影响,并提前消除非预期宕机等风险。

建设高可用架构

为了应对业务高并发及灵活的弹性容量管理需求,互联网行业在部署层面通常会采用分布式架构,因而比较容易构建单 ECS 实例容错能力。阿里云一个可用区一般是一套独立的 IDC 基础设施(机房供电、温控、物理网络等),可用区之间基础设施是隔离的,同地域的可用区之间通过内部专线网络打通。因此推荐以多可用区高可用架构为基础,进行单元化部署,合理规划业务单元容量,并制定相应的流量切换预案,通过一套“组合拳”提升应用系统单实例宕机容错和极端场景容灾能力。

1. 负载均衡 SLB 流量分发

业务实例挂载至 SLB 后端,通过 SLB 进行负载均衡和基于 ECS 实例健康检查的流量分发,提供单实例异常的容错能力,以及在 AZ 间进行流量切换的流量分发能力。

2. 单元化部署

业务采用多 AZ(可用区)形式部署,每一个 AZ 为一个独立的业务单元,具备全部业务功能。AZ 内部形成完整的上下游服务模块调用链,AZ 之间不存在模块的互相调用,避免形成 AZ 之间“拧麻花”式的调用关系耦合。

3. 容量规划 N+1 冗余

基于单元化部署进行容量规划,也就是如果有 N 个 AZ 则准备额外 1 个 AZ 的容量冗余并分摊至所有 AZ,以确保单 AZ 故障执行流量切换时的容量安全。

4. 具备应用系统流量切换能力

在前面的基础之上,应用层需建立 AZ 间流量切换能力,当单 AZ 下 ECS 实例批量故障或业务效果层面发生批量异常时,将单个异常 AZ 的流量切换至其他健康 AZ 进行止损。其优势在于,允许在大多数未定位故障根因的场景下作为通用预案来快速止损。

游戏行业最佳实践

游戏业务因为转化周期短、成本控制、合服滚服架构复杂度等因素，游戏服等核心业务很少会有自身架构上的单机高可用设计，业务稳定性强依赖云厂商，且对网络延迟、宕机敏感。尤其在新游戏上线初期，部分用户对游戏服 ECS 的稳定性甚至有着“0 宕机”的极致诉求。尽管 ECS 宕机后可以在秒级完成迁移并启动，但受游戏进程重新拉起、用户解密、加密磁盘等操作的耗时影响，宕机一般还是会造成游戏业务分钟级的不可用，以及内存数据未及时落盘导致的回档情况。且据调查结果显示至少有 60%-70% 的玩家反馈如果在游戏发布当天无法进行游戏，他们就不再有意愿接触这款游戏。所以游戏服 ECS 的稳定性保障是一个巨大挑战。阿里云经过众多游戏项目的沉淀打磨，形成了一套完善的游戏行业稳定性最佳实践。

通用最佳实践和互联网行业最佳实践中已详细介绍的内容不再赘述，以下将介绍游戏行业针对性的最佳实践。

选择合适的实例规格

游戏行业通常有几类典型业务场景，需要针对游戏业务特点选择合适的规格类型，常见的游戏业务 ECS 规格选型建议如下:

业务	特点	推荐选型
游戏服	单服高CPU，对计算性能和稳定性要求高	主售通用型实例，当前为：7代通用型规格c7\g7等
网关服	网络包转发、高吞吐，高网络PPS	主售网络增强型，当前为：7代网络增强型g7nex或6代增强型实例g6e
日志、中心服	分钟级异步入库、写频繁、高IOPS	选择ESSD磁盘规格，避免读写瓶颈

建立应用防抖能力

游戏服务器通常对性能抖动较为敏感，单实例防抖至关重要。建议结合业务的应用场景，对于实际应用中可能出现的短期抖动时长、响应时长进行统计和分位值分析，请求超时配置可以略大于业务可接受的响应时长分位值，避免单次请求失败率过高。同时，根据业务请求成功率要求，设定合理的重试次数与重试间隔时间。

选择合适部署方式

1. 使用部署集进行实例打散优化

相同用户的多个 ECS 实例部署在同一台物理服务器，会造成单台物理服务器宕机影响同用户的多个 ECS 实例的风险。通过部署集将 ECS 实例在物理服务器层面打散，可避免多台游戏服务器宕机造成大量玩家掉线，或避免数据库业务主、备库实例同时宕机而服务中断、无法快速恢复。建议用户将不可同时宕机的实例规划在一个部署集内，利用阿里云产品化能力保障重要实例之间的打散，降低“爆炸半径”。同时建议用户尽量分散可用区购买实例，减少对单一可用区的资源冲击，降低实例打散难度。

2. 使用专有宿主机进行部署优化

游戏场景通常有极致的性能和稳定性要求，诸如需要将一组游戏服和数据库实例放在同一台宿主机，进行高质量、低延迟的网络通信场景，可以通过专有宿主机进行自定义的部署规划来实现，并可自由决定主、备库实例在宿主机层面的打散，实现亲和与反亲和部署策略。

ECS 实例稳定性未来趋势

综上，阿里云 ECS 产品已经构建了完整成熟的稳定性体系，而未来还会继续在如下方向进行稳定性建设探索。

新场景

云原生、函数计算、Serverless PaaS 等场景下，“0 成本”获得更高的 ECS 稳定性体验；

智能化

可预测、可异构的无损热迁移和基于客户体感的智能灰度体系；

新一代“虚拟化”安全容器沙箱

软件定义的故障发现和隔离能力；

基础设施

OS+ 硬件 + 基础设施的一体化设计和定义。

通过稳定性体系的持续迭代演进，让客户在更丰富的场景下获得开箱即用的高稳定性、高性能和高性价比的 ECS 产品体验，让客户能够专注于自身业务系统的开发运营，助力客户业务发展。