容器服务

弹性推理服务

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

弹性推理服务
弹性推理服务产品简介

最近更新时间：2026-01-23 16:19:32

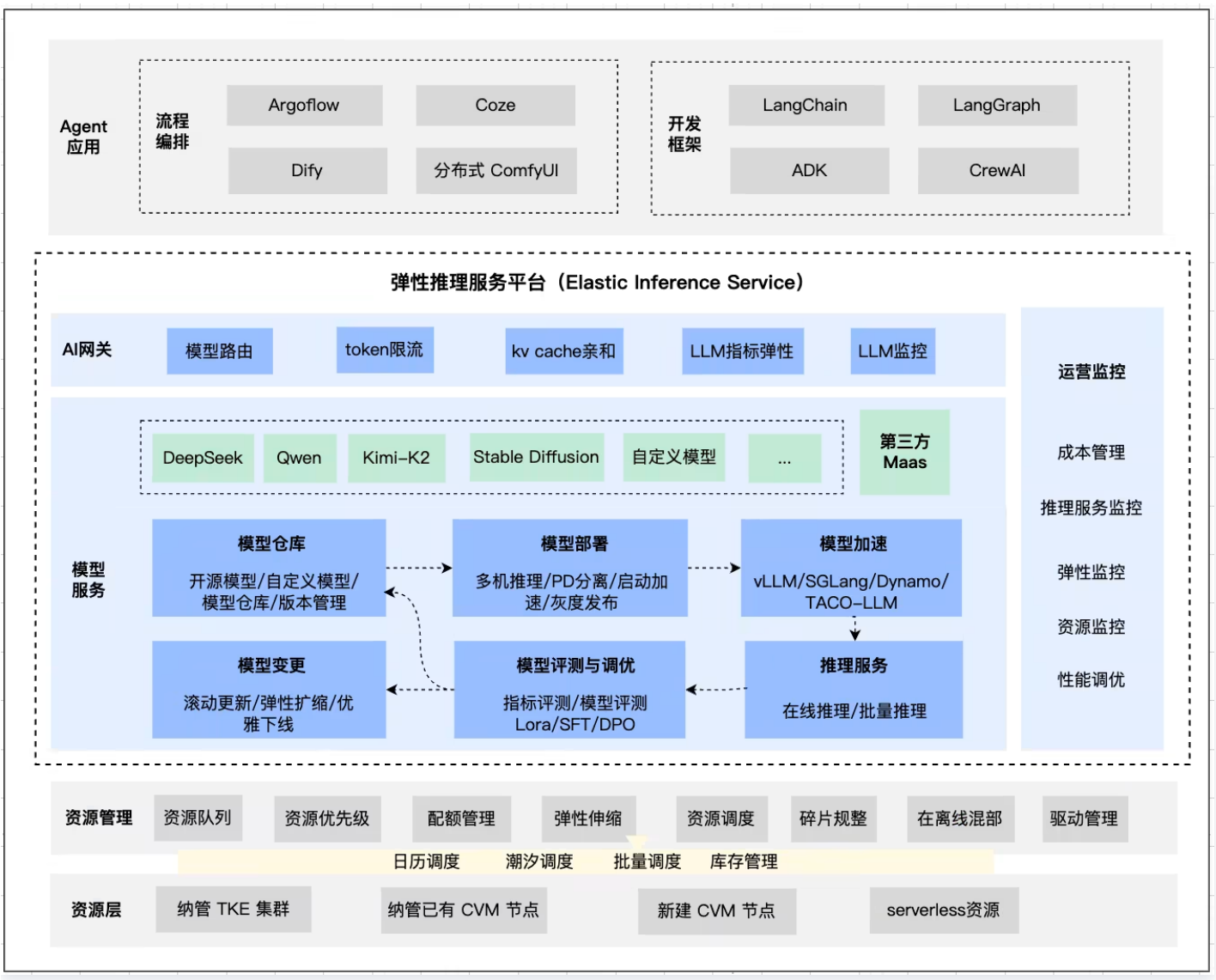
自建推理服务核心挑战

企业在自建并运维大模型推理服务的过程中，普遍会遇到四大核心挑战：

- **技术选型与适配复杂**：面对 vLLM、SGLang 等多样化的开源推理框架，企业不仅选型困难，后续的适配与优化工作也极为耗时。
- **GPU资源成本高昂**：推理业务普遍存在潮汐流量特性，推理波谷时段 GPU 只能空转，导致昂贵的 GPU 算力资源利用率低下，造成严重的成本浪费。
- **性能优化存在瓶颈**：大模型权重导致服务冷启动时间长，而针对特定硬件进行深度性能调优的技术壁垒高，难以达到理想的推理效率。
- **部署与运维难度高**：自行实现多机推理、PD 分离等高性能架构门槛极高，且后续的故障恢复、弹性伸缩、灰度发布等运维流程繁琐且易出错。

产品简介

弹性推理服务平台是基于腾讯云容器服务（TKE）构建的大模型推理服务平台，提供从模型部署、服务管理、推理加速到资源调度的一站式能力，帮助企业高效部署和管理生产级大模型推理服务。



应用场景

TKE 弹性推理服务 聚焦以下三大核心应用场景，帮助不同需求的企业快速落地大模型推理业务：

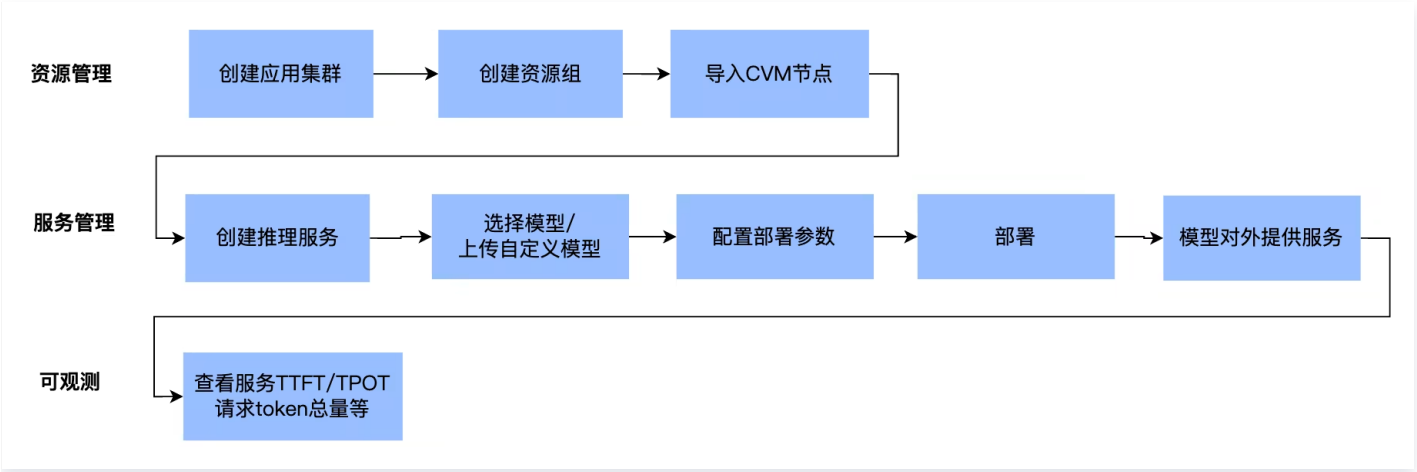
- **快速构建生产级 LLM 应用**：为寻求替代 MaaS、希望自主掌控 AI 能力的企业提供一站式解决方案。无论是大模型推理服务还是 AI+具体场景，用户都可一键部署经过深度优化的推理服务。
- **搭建企业内部 ML 平台**：作为一套 “AI Infra 套件”，TKE 弹性推理服务为企业提供搭建内部 ML 平台所需的核心底座。TKE 弹性推理服务和 TKE 生态深度集成，企业可将其无缝集成至自有工具链，使算法团队能聚焦于模型研发本身。
- **盘活存量 GPU 算力**：通过灵活的资源纳管方案，允许企业将已有的 CVM 统一接入 TKE 弹性推理服务进行调度，将闲置或存量算力高效用于 AI 推理任务，实现硬件成本的最大化利用。

核心优势

- **一键式部署与多框架兼容**：内置 vLLM、SGLang 等主流框架，并深度集成腾讯自研 TACO 加速框架。用户可通过控制台、CLI 或 API 从模型广场一键部署模型，彻底解决了“技术选型复杂”的痛点。

- **自研框架加速与架构优化**：通过集成 TACO 推理加速框架及原生支持 PD 分离架构，显著提升推理性能。内置的镜像与模型加载加速能力，可将服务冷启动时间缩短至秒级，有效突破了“性能优化瓶颈”。
- **企业级运维自动化**：平台完整接管了故障容错、模型路由、滚动更新、弹性伸缩等复杂的底层运维操作。这套自动化体系将团队从繁琐的部署运维工作中解放出来，确保推理服务的高可用性。
- **GPU 利用率最大化**：依托 TKE 弹性推理服务的资源混部与离线算力调度能力，能够智能地将潮汐流量下的闲置算力用于其他离线任务，有效拉升 GPU 综合利用率。
- **大模型推理专属监控**：提供从资源到服务的端到端可观测性，内置针对 TTFT（首字时延）、TPOT（ 输出吞吐率）等大模型核心指标的监控与告警，确保服务健康度的实时可观测。

使用流程



1. **准备推理资源**：首先，创建用于承载推理业务的应用集群。随后，在集群下创建资源组 ，并将您已购买或已有的 VM节 点导入该资源组，完成算力准备。
2. **部署推理服务**：在控制台单击“新建推理服务”，进入配置页面。从模型广场中选择目标模型，设定推理框架、部署架构（单机/多机/PD 分离）及服务访问方式后，即可一键部署。
3. **监控与日志**：服务成功运行后，利用平台集成的监控与日志功能，实时追踪服务的运行状态，并密切关注 TTFT、TPOT 等关键性能指标。
4. **服务生命周期管理**：在服务详情页，您可以对线上服务执行更新（支持滚动更新）、重启 、删除以及实例的扩缩容等全生命周期管理操作。

> ⚠ **注意：**
> 在进行资源准备时，请确保所选 CVM 节点的规格（尤其是 GPU 型号和显存）满足目标模型的推理要求，以避免部署失败。

相关服务

- TKE 弹性推理服务内的所有计算节点均由云服务器（CVM）实例（特别是 GPU 实例）提供。有关更多信息，请参见 云服务器产品文档 。

- TKE 弹性推理服务推理集群必须建立在私有网络（VPC）环境下，以保障网络的安全与隔离。集群内的所有节点和推理服务都在指定的 VPC 内进行通信。有关更多信息，请参见 私有网络产品文档 。
- 模型权重文件可以存放于对象存储（COS）中，并在创建推理服务时进行挂载，实现计算与存储的分离。有关更多信息，请参见 对象存储产品文档 。
- 当需要推理服务暴露至公网或内网进行访问时，TKE 弹性推理服务 会自动创建并绑定负载均衡（CLB）实例，以实现流量的分发和转发。有关更多信息，请参见 负载均衡产品文档 。
- TKE 弹性推理服务的监控数据可以对接到腾讯云 Prometheus 监控服务，日志数据可以投递至日志服务（CLS），实现对推理服务的统一观测和告警。有关更多信息，请参见 Prometheus 监控服务产品文档 和 日志服务产品文档 。

模型广场

最近更新时间：2026-01-23 16:19:32

概述

TKE 弹性推理服务模型广场针对模型部署时复杂的配置和调优问题，将业界主流的开源模型（如 DeepSeek、Qwen 等）与高性能推理引擎（如 TACO、vLLM 等）进行了预先集成和深度优化，让您可以跳过繁琐的准备工作，通过一键式操作快速部署获得稳定、强大的在线推理服务。

前提条件

在开始使用模型广场部署推理服务前，请确保您已满足以下条件：

1. 您已拥有腾讯云账号，并具备对相关云资源（如 COS、CLS 等）的操作权限。
2. 您已经按照 资源管理 的指引，创建了应用集群和资源组，并已导入可用的 CVM 计算节点。

操作步骤

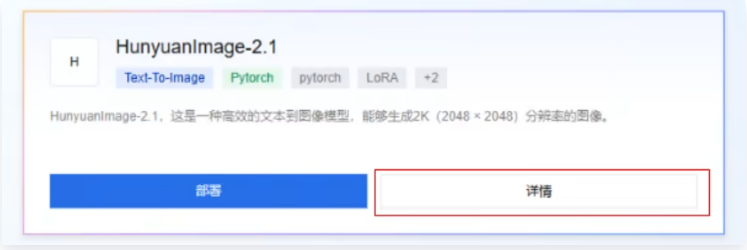
使用模型广场一键式部署推理服务

平台为您预置了多个业界主流的开源模型，无需您手动下载和配置，即可快速发起部署。

1. 登录 容器服务控制台，选择左侧导航栏中的**弹性推理服务**。
2. 在左侧导航栏中，单击**模型广场**，进入模型广场页面。 您可以在**模型广场**页面查看当前平台支持的所有内置模型列表，并通过任务类型、模型系列和模型相关标签进行筛选以便快速查找需要的模型。

> **说明：**
> 模型广场的内置模型列表由平台统一运营和更新，暂不支持用户在界面上进行自定义增删。

3. 单击**详情**，深入了解模型的全面信息。 这里是模型的数字档案，涵盖了其核心技术规格、性能指标和应用指南等内容（不同模型的涵盖内容会有差别）。

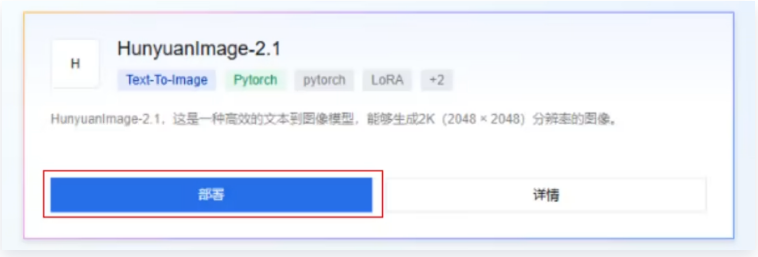


您将在这里找到：

- **核心信息**：模型版本、许可证、发布日期。
- **技术规格**：模型架构、参数量、所属框架（如 PyTorch、TensorFlow）。
- **模型能力**：适用任务、预期用途、性能评估（数据集与指标）。
- **使用指南**：快速上手的代码示例、API 调用方法。

> **说明：**
> 在部署前，请重点关注技术规格中的参数量，这直接决定了所需的 GPU 显存。例如，7B 模型通常需要至少 16GB 显存的 GPU。

4. 单击模型选项卡底部的**部署**，进入推理服务创建页面。



5. 在基本配置处，配置服务的基本信息。
   - **服务名称**：输入自定义的服务名称，例如 `my-hunyuan-service` 。
   - **描述**：根据需要填写服务的描述信息。
   - **资源模型**：在模型广场选择的模型，由平台自动设置。
6. 继续完成部署信息、访问配置 、高级参数等后续配置。 配置详情可以参考 推理服务 。
7. 确认所有配置信息无误后，单击底部**确定**，平台将开始为您部署所选模型的推理服务。服务部署通常需要3-5分钟，就绪后，您可以在推理服务列表中看到 my-hunyuan-service 的状态变为“运行中”。单击服务名称，即可在服务详情页面找到调用方法，以验证服务是否部署成功。

常见问题

我希望部署的模型不在模型广场列表中，应该怎么办？

TKE 弹性推理服务正在积极规划后续对其余模型的支持，欢迎向我们反馈您的模型需求，这将帮助我们确定未来优先支持的模型范围。 如果您已经拥有模型文件，也可以通过将模型文件上传至 COS 并进行授权，在 新建推理服务 时选择自定义模型以完成部署，具体步骤请参见 推理服务 。

选择不同模型时，对资源规格有什么建议？

不同参数规模的模型对 GPU 显存的要求差异很大。例如，Qwen/Qwen3-32B 这类百亿参数模型通常需要高端 GPU 才能运行。您可以在模型广场确认模型的参数大小并查看模型详情以获得具体的资源规格需求。

模型部署失败的常见原因有哪些？

除了资源规格不足外，常见的失败原因还包括：

- **网络配置问题**：请确保您的应用集群所在 VPC 网络通畅，特别是能够访问到模型所在的存储后端。
- **账户配额限制**：检查您的账户下相关资源（如 CVM、CLB）是否已达到配额上限。

相关文档

- 关于如何创建应用集群和资源组，并导入可用的 CVM 计算节点，请参见 资源管理 。

- 关于部署推理服务时的详细配置，请参见 推理服务 。

资源管理

最近更新时间：2026-01-23 16:19:32

概述

TKE 弹性推理服务专为模型推理服务建设了平台资源管理架构。通过资源管理功能，您可以将已有的 CVM 实例资源进行统一纳管，实现算力的集中。平台通过应用集群 – 资源组 – CVM 实例结构对资源进行组织，让您可以灵活地编排异构算力，并以极低的运维成本支撑上层推理业务的部署与生命周期管理。

核心概念

- **应用集群**：进行资源管理和推理服务部署的最高逻辑单元。其底层是一个经过平台封装的标准 Kubernetes 集群，无需进行额外运维操作。
- **资源组**：在应用集群内对 CVM 实例进行逻辑分组的单位，支持将不同规格和型号的异构 CVM 实例在一个集群下灵活编排。
- **CVM 实例**：提供计算能力的计算单元，由资源组统一管理和调度。

前提条件

在开始进行资源管理操作前，请确保您已满足以下条件：

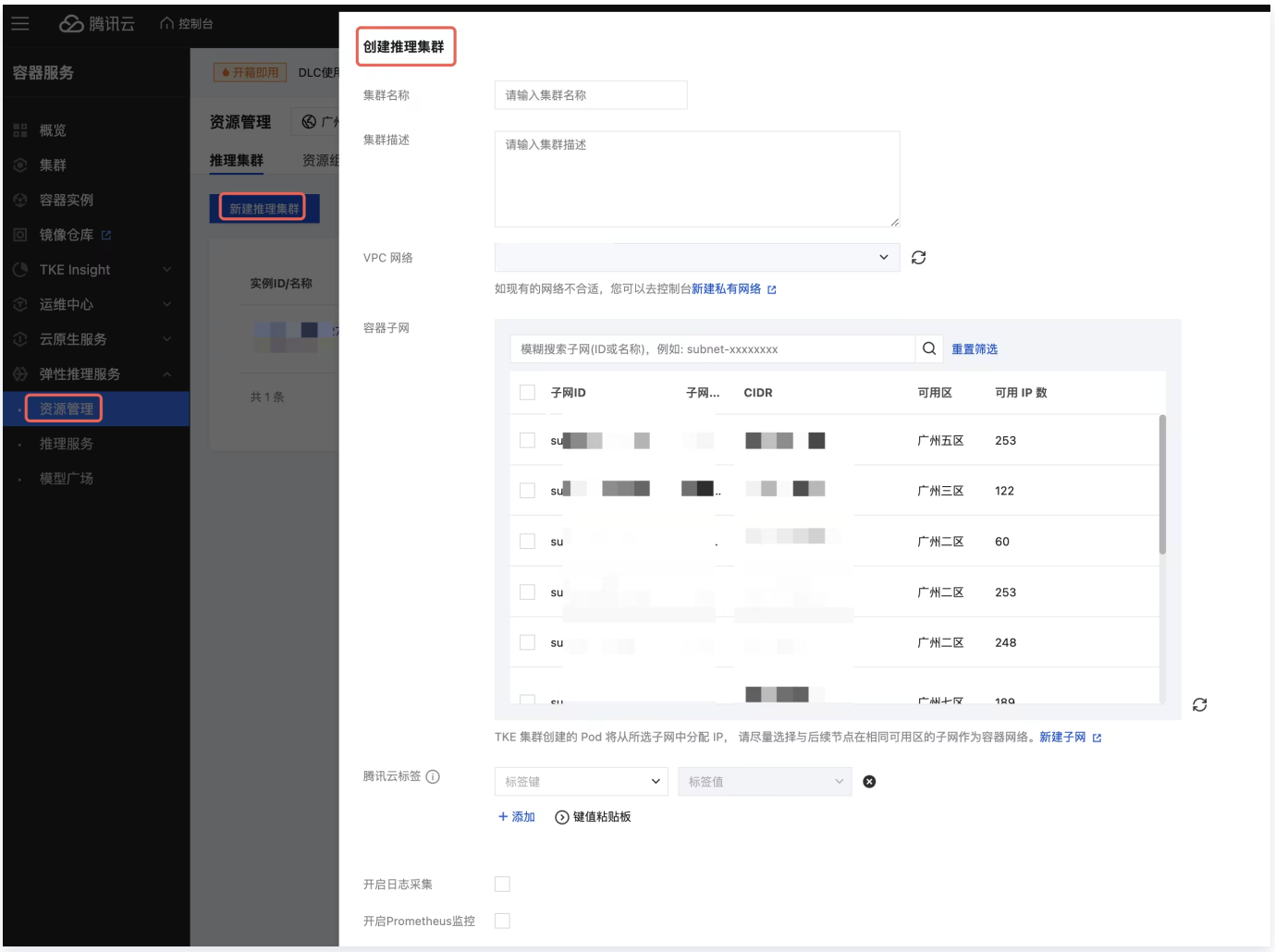
- 您已通过 内测申请问卷 填写申请并成功开通弹性推理服务平台。
- 您计划纳管的 CVM 实例已经存在，且与您计划创建的应用集群位于同一个私有网络（VPC）内。
- 您已拥有腾讯云账号，并具备对相关云资源（如集群、CVM 等）的操作权限。

操作步骤

一个典型的资源管理工作流包含**创建应用集群 > 创建资源组 > 导入 CVM 节点**三个主要步骤。

创建应用集群

1. 登录 容器服务控制台 ，选择左侧导航栏中的**弹性推理服务**。
2. 在左侧导航栏中，选择**资源管理**，进入推理集群列表页面。
3. 单击页面左上角的**新建推理集群**。
4. 在弹出的配置页面中，填写集群信息。



- **集群名称**：输入自定义的集群名称，例如 my-inference-cluster。
- **集群描述**：填写集群的描述信息，便于后续识别和管理，此项为选填。
- **VPC网络**：为集群选择一个合适的私有网络。

> **注意：**
> VPC 一旦选定后不可更改，且后续只能导入此 VPC 内的 CVM 实例，请谨慎规划您的网络。

- **容器子网**：为集群选择至少一个可用的子网。
- **腾讯云标签**：根据需要为集群添加标签。
- **日志采集**：开启后将自动推送推理服务日志到对应的日志集及日志主题，目前只支持配置已有日志集及日志主题。详情请参见 TKE 弹性推理服务可观测性 。
- **Prometheus 监控服务**：开启后，您可以按照实际需求灵活配置数据采集规则，其中基础指标永久免费提供监控，配置完成后即可在弹性推理服务平台查看监控数据，详情请参见 TKE 弹性推理服务可观测性 。

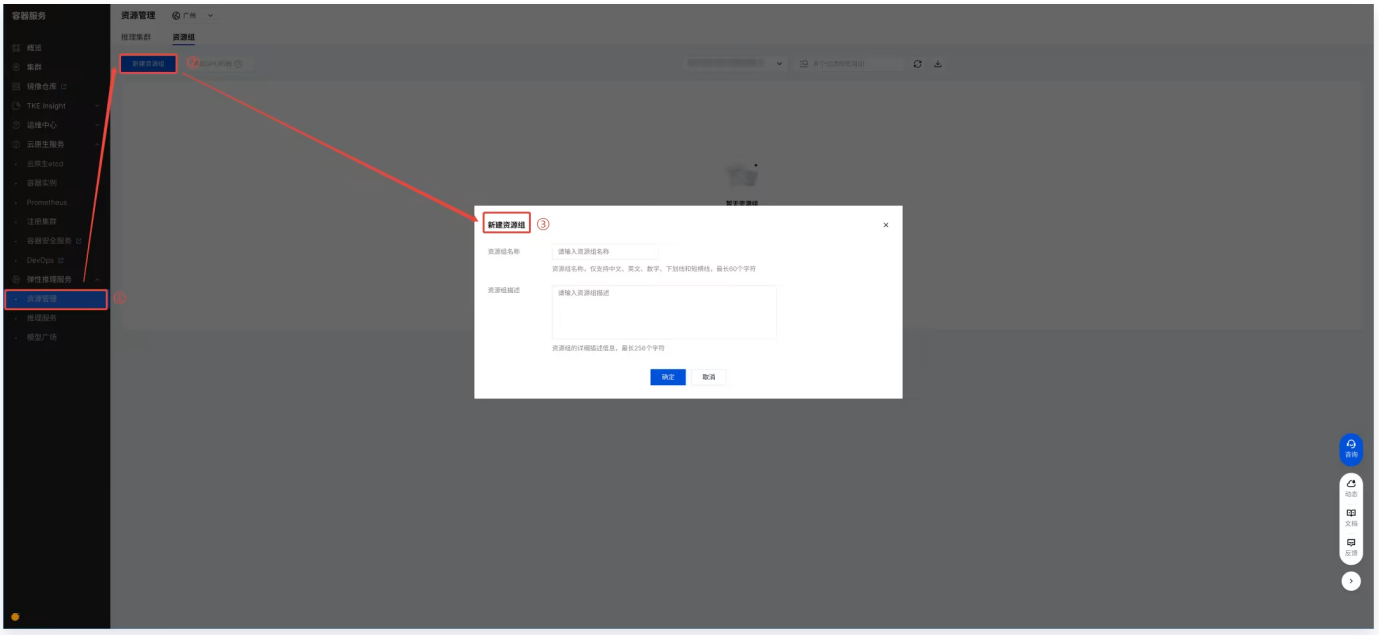
5. 确认配置无误后，单击**确定**。您将返回集群列表页面，并看到新创建的集群状态为“创建中”。

> **说明：**

> 默认情况下，每个账户最多支持创建5个应用集群。

创建资源组

1. 在**推理集群**列表中，单击您刚刚创建的集群 ID，进入其资源组列表页面。
2. 单击**新建资源组**。
3. 在弹出的窗口中，输入**资源组名称**和**备注**。
4. 单击**确定**，完成资源组的创建。



向资源组中导入 CVM 节点

> **警告：**
>
> **在导入 CVM 机器前，请您务必了解以下关键信息，以免造成数据丢失或管理问题：**
>
> - **数据将被清除**：为了确保环境的一致性，导入的 CVM 机器**需要重装操作系统，其系统盘上的所有数据都将被清除**。请在操作前务必做好**数据备份**。
> - **登录将被限制**：为实现统一管理并保障平台安全，导入弹性推理服务的 CVM 实例将限制用户通过常规方式直接登录。
> - **项目将自动归属**：导入操作完成后，CVM 实例的所属项目将自动变更为应用集群所指定的项目。

完成资源组创建后，您可以将已购买的 CVM 实例添加到组内，作为推理服务的算力资源。

1. 在**资源组**列表中，找到您需要操作的资源组，单击其右侧的**添加机器**。

2. 在弹出的机器列表中，勾选您希望导入的 CVM 实例。
   - 列表将自动筛选出与应用集群在同一个 VPC 下的所有 CVM 实例。
   - 已添加至其他集群的 CVM 实例不可重复添加。
3. 在添加 GPU 机器界面，参考以下提示进行配置：

4. 确认配置后，单击**添加到资源组**，开始导入过程。

后续操作

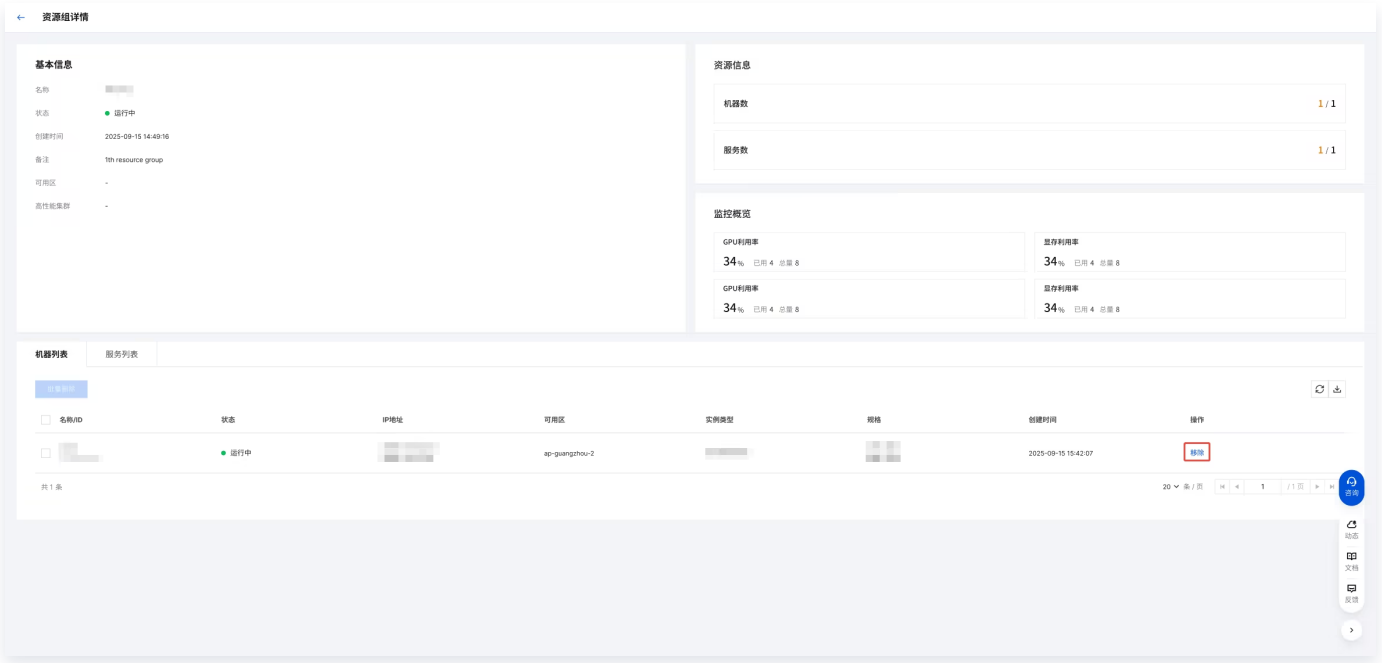
查看资源列表与详情

- 您可以在**推理集群**列表页查看所有集群的概览信息，包括集群 ID，运行状态、资源组数和机器数、创建时间等。
- 单击特定的集群 ID，可以查看该集群下的资源组列表，了解每个资源组的机器运行情况和少量监控信息。

- 单击特定的资源组 ID，可以进入**资源组详情**界面，查看该资源组的基本信息、监控信息、资源组下的机器列表以及部署在该资源组上的推理服务列表。

移除 CVM 实例

1. 在**推理集群**列表页单击集群 ID 进入**资源组列表页**。
2. 单击资源组名称，进入**资源组详情**界面。
3. 选择需要移除的 CVM 实例，单击列表中的**移除**。



4. 在确认弹窗中单击**确定**。被移除的机器将从弹性推理服务中解绑（但并不会销毁节点），并清除其上由弹性推理服务安装的相关组件。

删除资源组和应用集群

- **删除资源组**：在**资源组**列表中，找到您需要操作的资源组，单击其右侧 ⋮，在弹框中点击**删除**。**只有当资源组内没有任何 CVM 实例时**，才允许被删除，请务必在删除资源组前将其中的所有机器先行移除。

- **删除应用集群**：在**推理集群**列表中，找到您需要操作的集群，在右侧操作栏点击**删除**。**只有当应用集群内没有任何资源组和 CVM 实例时**，才允许被删除，请务必在删除集群前将其中的所有资源先行移除。

常见问题

为什么我需要先创建“应用集群”，再创建“资源组”？可以直接把 CVM 加到集群里吗?

不可以。TKE 弹性推理服务采用三层结构是为了实现更精细化的资源管理。**应用集群**定义了网络和安全边界，而**资源组**则是在此边界内对算力进行逻辑划分和调度的关键。这种设计允许您在同一个集群中安全地管理用于不同目的（如在线/离线）的异构资源，是实现资源高效利用和隔离的基础，因此资源组是必不可少的步骤。

导入 CVM节点时，提示“已添加至其他集群”是什么意思?

这意味着您尝试导入的 CVM 实例已经被另一个 Kubernetes 集群（无论是标准的 TKE 集群还是其他的应用集群）所管理。由于一个计算节点在同一时间只能被一个 Kubernetes 集群控制，因此您无法将其重复导入。请先将该节点从原集群中移除，然后再尝试导入弹性推理服务。

相关文档

- 关于Prometheus监控服务详情，请参见 Prometheus 监控概述 。
- 关于如何使用弹性推理服务的监控服务，请参见 TKE 弹性推理服务可观测性 。
- 关于如何使用应用集群和资源组部署推理服务，请参见 推理服务 。

推理服务

最近更新时间：2026-01-23 16:19:32

概述

TKE 弹性推理服务是模型发布为在线服务的核心组件，为开发者屏蔽了底层异构环境的复杂性，并提供了集成高性能引擎、支持多机多卡、PD 分离等高级部署架构的端到端推理能力。本文为您介绍如何使用弹性推理服务部署模型推理服务，并对其进行完整的生命周期管理。

前提条件

在创建推理服务前，请确保您已满足以下条件：

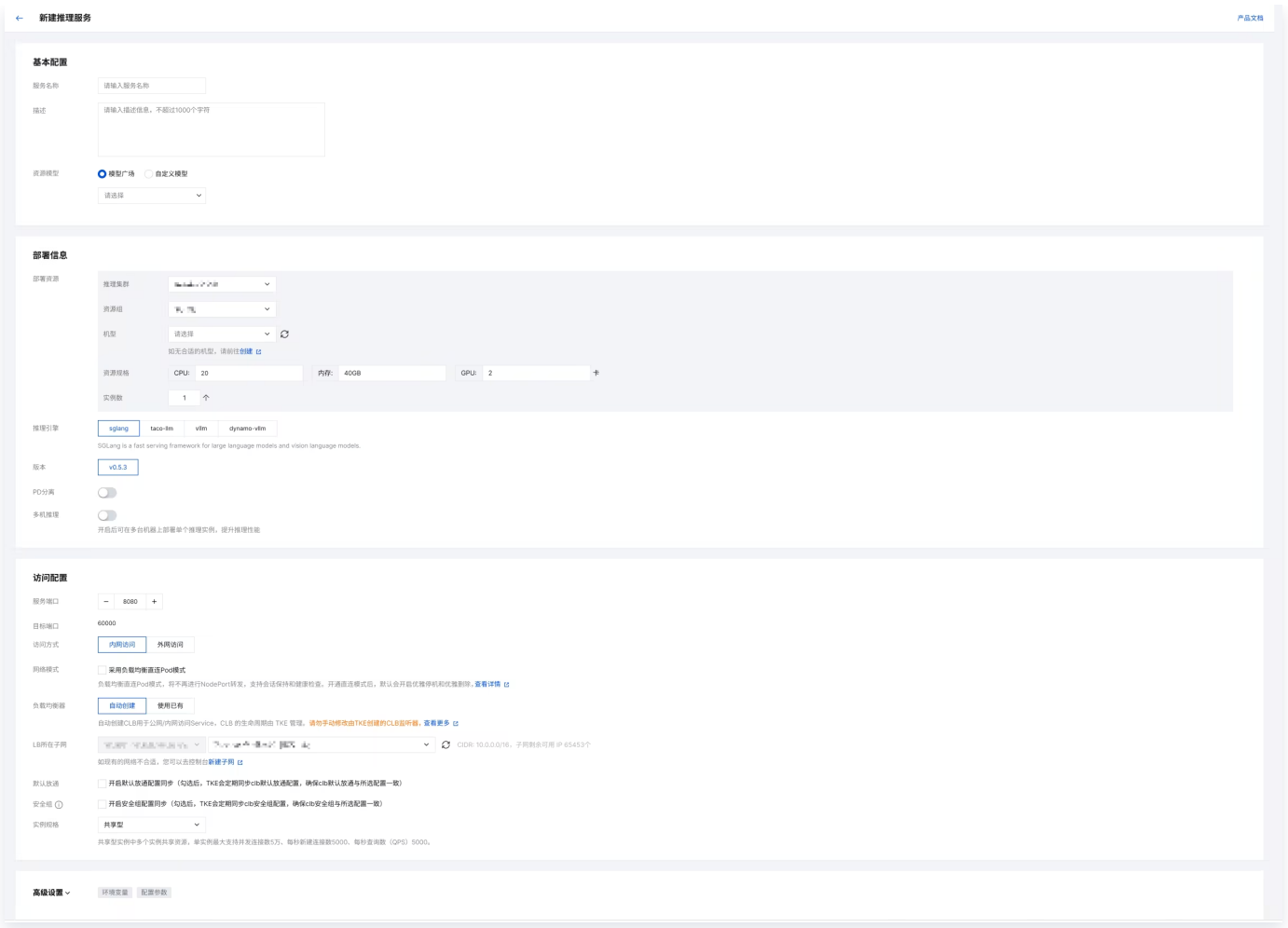
- 您已经成功开通弹性推理服务平台。
- 您已经按照 资源管理 的指引，创建了应用集群和资源组，并已导入可用的 CVM 计算节点。
- 您已拥有腾讯云账号，并具备对相关云资源（如 COS 等）的操作权限。

操作步骤

新建推理服务

新建服务是一个向导式的配置过程，引导您完成从模型选择到资源配置的全部步骤。

1. 登录 容器服务控制台 ，选择左侧导航栏中的**弹性推理服务**。
2. 在左侧导航栏中，单击**推理服务**，进入服务列表页面。
3. 单击页面左上角的**新建推理服务**。
4. 在新建服务页面，依次完成以下基本配置：



参数名称	参数说明
服务名称	输入一个在您账户下唯一的服务名称，用于标识该推理服务。
描述	填写服务的描述信息，便于后续识别和管理，此项为选填。
资源模型	选择用于推理服务的模型来源，可以选择平台提供的内置模型（参见 模型广场 ），或上传并选择 COS 路径以使用您自定义的模型（需要 COS 授权）。
部署资源	为服务指定运行的计算资源。您可以按序选择应用集群、资源组、机型三层计算资源选项，详情请参见 资源管理 。请确保所选资源组有足够的配额（CPU、内存、GPU）来满足后续实例规格和数量的需求。
推理引擎	根据您的模型和性能需求，选择合适的推理框架，弹性推理服务支持 vLLM、SGLang、TACO-vllm、Dynamo-vllm、 TACO-X。
版本	选择所使用推理引擎的版本。建议优先选择最新的稳定版本，以获得最佳性能和最新的功能支持。如果您的模型有特定的环境依赖，请选择与之兼容的版本。
PD 分离	此为针对大语言模型的高级性能优化选项。开启后，系统会将处理提示（Prefill）和生成内容（Decode）的计算任务分配到不同的实例组。适用于访问量大、请求中提示词较长、对响应延迟要求高的业务。

多机推理	适用于无法由单台机器承载的大模型。开启后，单个服务实例将跨多台机器进行分布式推理，用户可以手动指定一个实例部署的机器数。
服务端口与目标端口	定义流量访问的端口映射。 • **服务端口**：是负载均衡器（CLB）上暴露给外部访问的端口，例如8080。 • **目标端口**：是运行模型的容器内部实际监听的端口，弹性推理服务限定为60000。
访问方式	配置服务的网络可访问性。 • **内网访问**：服务只能在同一VPC（私有网络）内被调用，安全性高，适用于集群内部微服务调用场景。 • **公网访问**：平台将自动创建并绑定一个公网CLB，生成一个公网地址，允许从互联网直接访问服务，适用于对外提供API的场景。
网络模式	支持采用负载均衡直连 Pod 模式，勾选后会自动开启优雅停机和优雅删除。详情请参见 在 TKE 上使用负载均衡直连 Pod 。
负载均衡器	为服务配置流量入口。可选择新建 CLB，由平台自动创建并关联；或选择绑定到账户下已有的 CLB 实例，以复用现有资源和访问入口。
LB所在子网	为负载均衡器（CLB）实例指定一个部署的子网。请确保所选子网有足够的可用 IP 地址，并规划好网络访问控制（如与后端服务的网络互通性）。
默认放通	持续同步后端服务器的安全组，以自动放通来自 CLB 的访问流量，确保 CLB 到服务的网络路径始终畅通。
安全组	持续同步您为 CLB 实例本身绑定的安全组，以确保服务对公网的访问权限始终与您在弹性推理服务中的配置保持一致。
实例规格	选择实例的性能类型，这将决定其资源隔离级别和网络转发性能。 • **共享型**：实例间共享资源，适用于开发测试或流量较小的业务场景。单实例最大支持并发连接数5万、每秒新建连接数5000、每秒查询数（QPS）5000。 • **性能容量型**：独享转发性能，不受其他实例影响，适用于生产环境或对性能、并发有高要求的业务场景。单实例最大支持并发连接数10,000,000，新建连接数1,000,000，每秒查询数300,000。

5. （可选）您可以展开**高级配置**，对所选推理框架的参数进行精细化调整，具体参数如下表：

说明：

根据在前面流程中选择的不同的推理引擎，此环节的参数配置会有所不同。

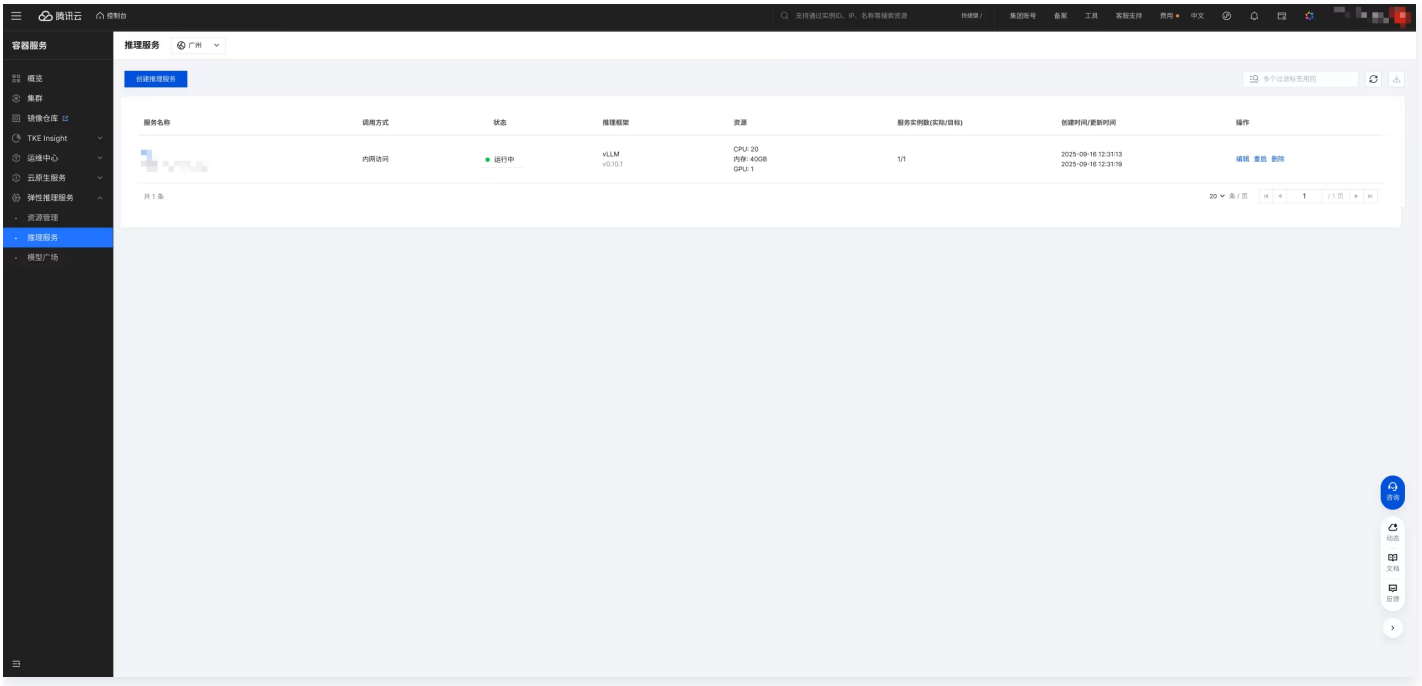
参数名称	参数说明
环境变量	类型：string 说明：以 key=value 的形式为服务容器注入自定义环境变量。
最大并发处理数(maxBatchSize)	类型：int 说明：推理引擎在单个 batch 中能并行处理的最大请求数量。此值影响服务的并发能力和吞吐量，需根据显存大小进行权衡。
最大上下文长度(maxModelLen)	类型：int 说明：定义模型能处理的最大序列长度。此值受限于模型的预训练设置和 GPU 显存大小，设置过高可能导致显存溢出。
张量并行数(tpSize)	类型：int 说明：将模型的权重沿特定维度切分到多个 GPU 上，以减少单个 GPU 的显存占用。是解决大模型单卡无法加载问题的常用方法。
流水线并行数(ppSize)	类型：int 说明：将模型的不同 Layers 分布到不同的 GPU 上，数据在 GPU 间以流水线方式处理。适用于超大规模模型的训练和推理。 约束：通常与 tp_size 结合使用，tp_size * pp_size 等于您为单个副本分配的总卡数。
数据并行数(dataParallelSize)	类型：int 说明：将相同的模型副本部署在多个 GPU 上，并将输入数据 batch 切分后并行处理。主要用于提升吞吐量。
显存利用率(gpuMemoryU	类型：float

tilization)	说明：设置推理服务可以使用的单张 GPU 显存的比例上限。这个参数用于为 KV Cache 预留空间。KV Cache 的大小是动态的，取决于批次大小和序列长度。示例：gpuMemoryUtilization为 0.85 意味着服务会预先分配 85% 的显存用于加载模型权重和计算，剩余的 15% 留作他用或作为安全缓冲。这有助于防止因 KV Cache 增长而导致的显存溢出。
单批次最大Token数(maxNumBatchedTokens)	类型：int 说明：定义了推理引擎在一个批次中能够处理的所有序列的 Token 总数的上限。
最大并发处理数(maxNumSeqs)	taco 专属参数，意义同 maxBatchSize。
流水线并行数(pipelineParallelSize)	taco 专属参数，意义同 ppSize。
张量并行数(tensorParallelSize)	taco 专属参数，意义同 tpSize。

6. 确认无误后，单击**确定**，平台将开始部署您的推理服务。

查看服务状态与详情

服务创建后，您将在**推理服务**列表页看到新创建的服务及其状态。



- **可查看内容**：推理服务名、调用方式、运行状态、推理引擎/框架、资源、实例数、创建/更新时间。

- **运行状态说明**：
  - **等待中**：服务正在部署过程中，如拉取镜像、加载模型等。
  - **运行中**：服务已成功启动并通过健康检查，可以正常接收请求。
  - **异常**：服务实例出现错误，无法正常运行。

单击服务名称，即可进入**服务详情**页面。在此页面，您可以查看：

- **基本信息**：服务的完整配置。
- **实例列表**：服务下所有实例的运行状态、所在节点IP等。
- **访问方式**：服务的内网和公网访问地址及调用示例。
- **监控**：服务的核心性能指标图表，如首字延迟 (TTFT)、字间延迟 (TPOT) 和端到端 (E2E) 请求延迟等。
- **日志**：查看服务实例的实时日志。
- **YAML**：查看该服务在底层 Kubernetes 中的资源定义文件。

管理服务生命周期

在**推理服务**列表页或详情页，您可以对服务进行以下管理操作：

- **更新服务**：单击**编辑**，您可以修改服务的资源规格、实例数、弹性配置、高级参数等。

> **说明：**
> 推理服务的推理框架、部署架构等基础配置在创建后不可变更。

- **重启服务**：单击**重启**，平台将重新部署服务的所有实例。
- **删除服务**：单击**删除**，平台将执行优雅下线，并在终止服务前处理完所有进行中的请求，然后释放相关资源。

相关文档

- 关于如何创建应用集群和资源组，并导入可用的 CVM 计算节点，请参见 资源管理 。
- 关于内置的模型列表，请参见 模型广场 。

可观测性

最近更新时间：2026-01-23 16:19:32

概述

弹性推理服务平台提供开箱即用的可观测性能力，帮助您全面洞察推理服务的运行状态。您可以在控制台通过可视化的图表，实时监控服务的核心性能指标与资源使用情况，并结合日志功能快速诊断和定位问题。

前提条件

在使用 EIS 可观测性功能前，请确保您已满足以下条件：

- 您已经成功开通弹性推理服务平台，并已部署了至少一个推理服务。
- 您已经成功开通腾讯云日志服务（CLS），并创建了用于接收推理日志的日志集和日志主题。
- 如果您希望将监控数据对接到自有的监控体系，请确保您已部署并运行了 Prometheus 服务。

监控

EIS 监控体系遵循自顶向下的问题排查思路，为您提供了从推理服务、节点到应用集群的三个层级监控视图，以满足业务健康度巡检、性能瓶颈定位和资源容量规划等不同运维场景的需求。

推理服务监控视图

用于业务健康度巡检。这是日常监控的主要入口，用于从业务视角快速评估服务的整体性能表现和请求处理情况。

操作路径：

1. 登录 容器服务控制台 ，进入**弹性推理服务** > **推理服务**列表。
2. 单击目标服务的名称，进入其**服务详情**页面。
3. 选择**监控**页签即可查看。

监控指标：

- **服务性能指标**：主要关注服务响应速度和内部运行状态。响应速度指标包括 **TTFT (首字延迟)**、**TPOT (字间延迟)** 和 **E2E (端到端) 请求延迟**；内部状态指标包括**调度器中运行和等待的请求数**以及**GPU KV 缓存使用率**。除了上述默认指标以外，您可以手动添加更多指标，详细指标清单参见 vLLM 监控指标 、SGLang 监控指标 与 Dynamo 监控指标 。
- **服务资源指标**：服务下所有实例平均的 GPU 使用率、GPU 显存使用量、GPU 显存使用率、CPU 使用率和内存使用率等。

节点监控视图

用于性能瓶颈定位。当服务指标出现异常时，可以下钻到此视图来分析具体节点的资源消耗，定位问题根源。

操作路径：

1. 在**服务详情**页面，选择**实例列表**页签。
2. 单击目标服务实例所在**节点**的 ID，即可跳转至该节点的监控详情页。

监控指标：

- **服务资源指标**：该节点实时的 GPU 使用率、GPU 显存使用量、GPU 显存使用率、CPU 使用率和内存使用率等详细资源指标。

应用集群监控视图

用于资源容量规划。此视图可以评估整个应用集群的健康度和容量水位，为扩缩容等规划活动提供数据支持。

操作路径：

1. 在左侧导航栏中，单击**应用集群**，进入集群列表页面。
2. 单击目标集群的名称，进入其**集群详情**页面。
3. 选择**监控**页签即可查看。

监控指标：

- **服务资源指标**：集群内所有节点聚合后的平均 GPU 使用率、GPU 显存使用量、GPU 显存使用率、CPU 使用率和内存使用率等指标。

对接自有 Prometheus

在**推理服务**页面，您可以配置将监控指标投递至您自有的 Prometheus 服务，以便集成到统一的告警体系中。

1. 打开**推理服务**页面的**监控**部分。
2. 开启**监控**功能，并填写您的 Prometheus 服务地址和相关认证信息。

日志

EIS 平台支持将服务日志自动投递至腾讯云日志服务（CLS），并提供了专为 AI 应用场景设计的结构化日志检索能力。

日志采集与投递

- **标准容器日志**：平台默认采集并投递服务容器的标准输出和标准错误日志流。
- **AI 应用日志**：平台支持对结构化日志的解析。通过在日志中包含特定字段，您可以启用按请求或会话维度的日志检索，极大提升 AI 应用的问题排查效率。

操作步骤

配置日志采集

您可以在创建推理服务时，为其配置日志采集规则。

1. 在**新建服务**或**更新服务配置**页面的**日志**部分开启日志功能。
2. **选择日志集**：从下拉列表中选择一个您在 CLS 中预先创建好的日志集。
3. **选择日志主题**：选择该日志集下的一个日志主题作为日志投递的目标。

查看与分析日志

1. 进入目标服务的**服务详情**页面。

2. 选择**日志**页签。
3. 在此页面，您可以实时查看服务下所有实例的日志流。您可以通过实例（Pod）筛选、关键词搜索等方式快速定位您关心的日志内容。
4. 如果需要进行更高级的检索分析，您可以在日志服务（CLS）中进行深度分析。

相关文档

- 关于完整的资源指标列表，请参见 Prometheus 监控服务 容器监控图表指标 。
- 关于 vLLM 框架支持的推理监控指标，请参见 vLLM Metrics 。
- 关于 SGLang 支持的推理监控指标，请参见 SGLang Production Metrics 。
- 关于 Dynamo 支持的推理监控指标，请参见 Dynamo MetricsRegistry 。