弹性内存存储

产品介绍

文档版本 01

发布日期 2025-09-15

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 什么是弹性内存存储 1
2 产品优势 3
3 应用场景 4
4 产品功能 6
5 EMS 以存代算 7
6 安全 9
6.1 责任共担 9
6.2 身份认证与访问控制 10
6.3 数据保护技术 11
6.4 认证证书 12
7 计费说明 14
8 约束与限制 16
9 与其他服务的关系 17
10 基本概念 19
10.1 EMS 基本概念 19

1 什么是弹性内存存储

弹性内存存储（Elastic Memory Service，EMS）是一种以DRAM内存（动态随机存取存储器）为主要存储介质的云基础设施服务，为LLM推理提供缓存和推理加速。EMS实现AI服务器的分布式内存池化管理，将LLM推理场景下多轮对话及公共前缀等历史KVCache缓存到EMS内存存储中，通过以存代算，减少了冗余计算，提升推理吞吐量，大幅节省AI推理算力资源，同时可降低推理首Token时延（Time To First Token，TTFT），提升LLM推理对话体验。

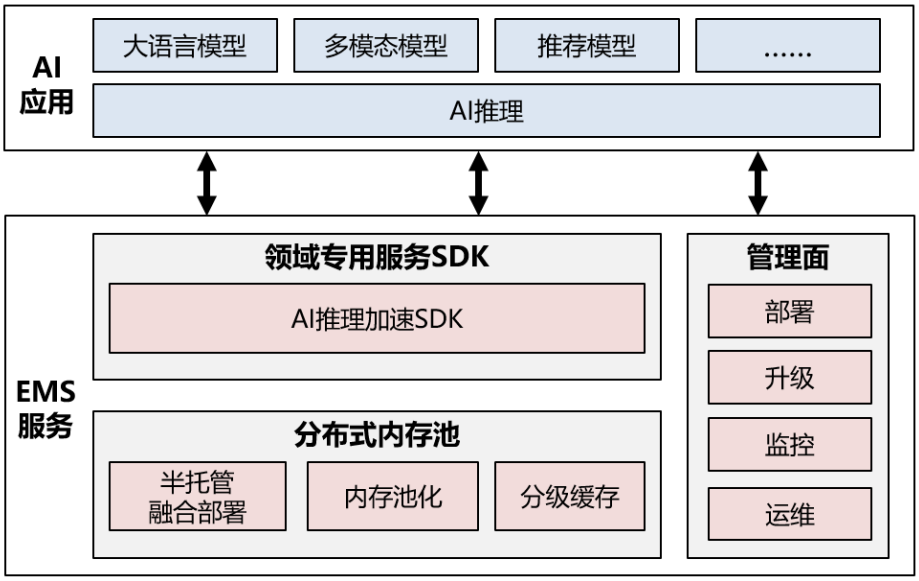
通过EMS，华为云将传统的“计算-存储”分离的两层云架构升级为“计算-内存-存储”的三层云架构，其中新增的“内存层”即为弹性内存存储。这种新型的三层云架构能有效解决存力痛点，从而具有高资源弹性、高资源利用率和高性能等优势。具体来说，EMS通过以下方式解决AI时代的存力问题：

- **提升数据访问速度：**针对AI场景中“持久化存储性能不足”的问题，EMS作为计算层与存储层之间的高性能内存缓存层，利用DRAM内存来缓存持久化存储层的数据或在计算过程中产生的中间数据。
- **高效利用DRAM资源：**针对AI场景中“DRAM内存利用率较低”的问题，EMS将AI服务器中的空闲DRAM资源进行池化，形成EMS内存池，实现DRAM资源的按需分配和高效利用。
- **提升AI推理性能：**针对AI推理场景中的“显存内存墙”问题，EMS利用内存池中的DRAM资源进行扩展，通过DRAM内存容量和带宽的补充，大幅提升AI推理的性能。

产品架构

EMS产品架构主要由三部分组成：领域专用服务SDK、分布式内存池和管理面。请参考**图1 EMS产品架构**。

图 1-1 EMS 产品架构



- **领域专用服务SDK**包含一系列面向不同AI应用场景的插件和接口服务SDK，提供业务系统接入、业务数据布局和近数据处理等功能，实现业务请求的内存加速。目前，该SDK主要应用于大语言模型的推理，通过分布式内存池提升处理效率并降低成本。
- **分布式内存池**负责跨节点的内存空间管理、数据负载均衡等任务，通过空间池化提供内存缓存共享访问。内存池当前采用融合部署方式，即利用AI服务器中的DRAM，将DRAM内存池化以实现分布式共享，并进行本地亲和调度和访问。
- **EMS管理面**负责EMS服务的部署、监控、升级及运维管理等功能，通过华为云的云原生基础设施为用户提供一站式的云上运维解决方案。

访问方式

关于EMS资源发放部署等操作，请使用控制台方式访问弹性内存存储服务。

基于控制台方式和SDK方式访问弹性内存存储服务：

- 控制台方式

关于EMS资源发放部署等操作，请使用控制台方式访问弹性内存存储服务。

- SDK方式

模型推理框架（如：vLLM）及企业自研的推理框架通过集成EMS SDK方式访问弹性内存存储服务，具体操作请参见弹性内存存储服务SDK参考。

2 产品优势

EMS内存存储具有以下优势：

- **半托管融合部署，降低成本**

  EMS数据面部署在AI服务器上，采用融合部署，统一纳管AI服务器上空闲的DRAM内存资源，复用DRAM内存资源，提供推理加速服务，降低推理KVCache存储成本。

- **分级缓存，提升推理吞吐，优化推理时延**

  EMS通过构建“显存-内存-存储”三级缓存体系，实现历史KVCache动态分层存储，突破显存瓶颈，实现显存扩展；通过缓存推理历史KVCache，实现以存代算，提升LLM推理服务的吞吐性能，降低推理资源成本；同时缩短LLM推理首Token输出时延，改善用户对话体验。

- **分布式共享内存池，提升缓存命中率**

  EMS将AI服务器上空闲的DRAM内存构建成分布式内存池，突破单机内存瓶颈，提升缓存空间，同时使得节点间能够进行高效的数据共享，支持亲和调度，提升缓存命中率，满足大规模分布式推理需求。

- **兼容主流推理框架，满足多样化访问**

  EMS提供SDK，供各种推理框架集成，兼容vLLM等开源框架及其他企业自研的LLM框架，适配LLM推理环节中对内存Cache的多样化访问需求。

3 应用场景

LLM 大语言模型推理

需求和挑战

随着LLM推理的飞速发展，LLM推理需求急速增加，LLM推理包含多种任务，如：多轮对话交互、信息检索和文本生成（包括代码）等。

LLM推理场景的需求和主要挑战如下：

- **保持连贯性：**受限于显存容量原因，多轮交互使智能助手很容易“忘记”对话中更早的部分或重复自己说过的话。
- **推理吞吐性能低：**LLM在线推理需要满足大量消费者用户同时使用，受限于AI显存内存墙瓶颈，单卡推理吞吐性能低，大量用户并发访问时时延高，导致用户需要部署大量AI推理算力资源，推理吞吐资源成本高。
- **推理延迟高：**在大模型推理过程中，从输入指令到模型产生预测并输出内容的时间过高，严重影响用户体验，尤其是和智能助手进行多轮对话时。

解决方案

针对AI推理场景面临的痛点问题，华为云通过EMS加速推理业务，提升推理业务吞吐，降低推理时延，降低推理资源部署成本。LLM在线推理场景示意图如图3-1所示。

由于AI服务器显存内存墙瓶颈，EMS利用AI集群的空闲内存构建分布式内存池，实现显存容量的扩展，突破单机内存的瓶颈。通过EMS将LLM推理中的多轮对话、公共前缀等场景下的历史KVCache缓存在EMS中，LLM推理时直接复用EMS缓存中的历史KVCache，无需重新计算历史KVCache，通过以存代算，降低了推理首Token时延（Time To First Token，TTFT），同时也节省了推理算力，提高推理吞吐，加速了大模型推理服务的效率。

建议搭配服务

AI开发平台 ModelArts、云容器引擎 CCE、高性能弹性文件服务 SFS Turbo、对象存储服务 OBS。

图 3-1 LLM 大语言模型推理

4 产品功能

表4-1列出了弹性内存存储服务EMS提供的常用功能特性。

表 4-1 EMS 功能概览

功能名称	功能描述
创建凭证	使用EMS前，需要先创建凭证，用于激活EMS。
部署EMS	1. 在已创建的CCE集群的节点上部署EMS，以提供内存服务。 2. 在CCE集群上安装监控插件，将CCE集群上部署的EMS监控数据上报至AOM实例，便于您随时监控业务。 3. 在CCE集群上配置告警规则，出现EMS告警时，能够及时通知您处理告警。 4. 通过将EMS日志规则配置到云日志服务 LTS，您可以获取EMS的相关操作日志，从而帮助您定位问题。
激活EMS	使用已创建的凭证激活EMS后，才能正常使用EMS。
使用EMS	您的推理框架（如：vLLM）可以通过集成EMS SDK方式访问EMS弹性内存存储服务，以实现推理KVCache缓存及后续访问命中。
升级EMS	如果您的EMS软件版本较低，可以执行EMS升级。
卸载EMS	如果您的业务不再使用EMS，可以卸载EMS。

⚠ 注意

激活凭证将作为软件license关联您的EMS软件使用计费，请您妥善保管激活凭证，避免泄露。

5 EMS 以存代算

以存代算产生的背景

在AI推理过程中，Transformer模型接收用户的问题输入，并通过迭代方式生成相应的回答。每个Transformer层由自注意力模块和前馈网络模块组成。

在自注意力模块中，上下文词元（token）与模型参数结合，生成中间数据K（键）和V（值），并进行注意力计算。为避免在迭代生成过程中重复计算KV，生成的KV中间数据被存储在AI服务器的显存中，形成KV缓存。每个词元的KV缓存大小取决于模型的维度、层数以及数据精度，计算公式为：单个词元的KV缓存大小 = 模型维度 * 模型层数 * 数据精度 * 2。例如：GPT3模型的数据维度和层数分别为12288和96，在双字节精度下，单个词元的KV缓存大小为12288 * 96 * 2 * 2字节 = 4.5MB。

在推理过程中，每个推理请求所需的KV缓存大小与上下文长度成线性关系。例如：在GPT3模型的推理中，长度为2048的上下文将占用约4.5MB * 2048 = 10GB的AI服务器显存空间。

然而，AI服务器通常只能提供几十GB的显存容量，其中一部分还要用于存储模型参数，仅剩余部分空间用于KVCache缓存。例如：使用8张64GB的AI服务器部署GPT3模型，系统显存总容量为512GB（8 * 64GB），其中350GB用于模型参数，剩余162GB仅能支持16个（162GB / 10GB）2048上下文长度的推理请求缓存KV值。

因此，AI服务器能够同时处理的请求数量受限于显存容量。

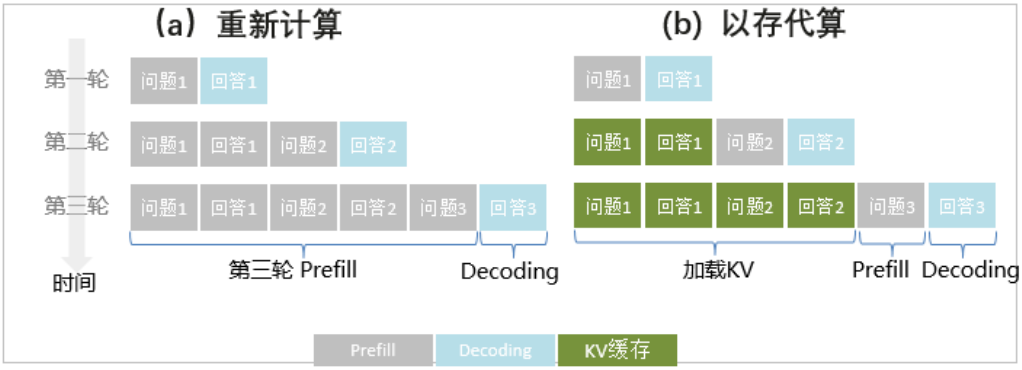
综上所述，Transformer模型推理中存在严重的AI显存内存墙问题。为解决这一问题，EMS通过以存代算技术加速AI推理。

以存代算原理

在Transformer模型的推理过程中，由于AI服务器的显存容量限制，现有的推理系统无法在AI服务器的显存中持续保存多轮对话的KVCache缓存。为了应对这一问题，系统通常会丢弃已处理对话的KV缓存，以腾出显存空间来服务新的请求。然而，当这些被丢弃的KV缓存对应的对话再次出现时，系统必须重新计算这些KV缓存，如图5-1 中的（a）所示。这种重复计算不仅浪费了计算资源，还增加了推理成本。

为了减少成本并提升推理性能，EMS服务引入了以存代算技术CachedAttention，如图5-1 中的（b）所示。该技术利用EMS中的大容量分布式内存池来存储和复用多轮对话中产生的KVCache缓存，而不是直接丢弃它们。具体操作是，将一个会话对应的历史KV缓存保存到EMS中，当该对话重新激活时，再从EMS中加载并复用这些KV缓存，从而避免了重复计算。

图 5-1 多轮对话中使用 EMS

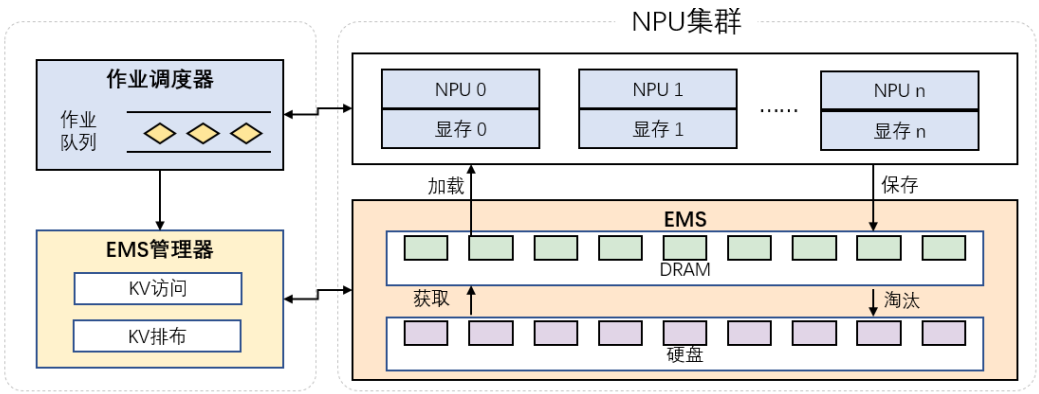


通过以存代算技术，EMS有效地避免了多轮对话中的重复计算，显著降低了首字时延，提高了推理预填充阶段（Prefill阶段）的吞吐量，并降低了端到端的推理成本。

缓存系统性能优化

EMS还采用了以下技术来优化缓存系统性能，如图5-2所示：

图 5-2 EMS 以存代算技术



- 采用异步保存和读取策略，以减少加载和保存KV缓存的时间。
- 利用多级缓存和分布式缓存架构，通过更大容量的存储介质提供充足的缓存空间。
- 通过自动感知调度器中的任务队列信息，实现多层次存储介质间的缓存调度，以提高访问效率。

6 安全

6.1 责任共担

华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

与传统的本地数据中心相比，云计算的运营方和使用方分离，提供了更好的灵活性和控制力，有效降低了客户的运营负担。正因如此，云的安全性无法由一方完全承担，云安全工作需要华为云与您共同努力，如图6-1所示。

- **华为云**：无论在任何云服务类别下，华为云都会承担基础设施的安全责任，包括安全性、合规性。该基础设施由华为云提供的物理数据中心（计算、存储、网络等）、虚拟化平台及云服务组成。在PaaS、SaaS场景下，华为云也会基于控制原则承担所提供服务或组件的安全配置、漏洞修复、安全防护和入侵检测等职责。
- **客户**：无论在任何云服务类别下，客户数据资产的所有权和控制权都不会转移。在未经授权的情况，华为云承诺不触碰客户数据，客户的内容数据、身份和权限都需要客户自身看护，这包括确保云上内容的合法合规，使用安全的凭证（如强口令、多因子认证）并妥善管理，同时监控内容安全事件和账号异常行为并及时响应。

图 6-1 华为云安全责任共担模型

云安全责任基于控制权，以可见、可用作为前提。在客户上云的过程中，资产（例如设备、硬件、软件、介质、虚拟机、操作系统、数据等）由客户完全控制向客户与华为云共同控制转变，这也就意味着客户需要承担的责任取决于客户所选取的云服务。如图6-1所示，客户可以基于自身的业务需求选择不同的云服务类别（例如IaaS、PaaS、SaaS服务）。不同的云服务类别中，每个组件的控制权不同，这也导致了华为云与客户的责任关系不同。

- 在On-prem场景下，由于客户享有对硬件、软件和数据等资产的全部控制权，因此客户应当对所有组件的安全性负责。
- 在IaaS场景下，客户控制着除基础设施外的所有组件，因此客户需要做好除基础设施外的所有组件的安全工作，例如应用自身的合法合规性、开发设计安全，以及相关组件（如中间件、数据库和操作系统）的漏洞修复、配置安全、安全防护方案等。
- 在PaaS场景下，客户除了对自身部署的应用负责，也要做好自身控制的中间件、数据库、网络控制的安全配置和策略工作。
- 在SaaS场景下，客户对客户内容、账号和权限具有控制权，客户需要做好自身内容的保护以及合法合规、账号和权限的配置和保护等。

6.2 身份认证与访问控制

IAM身份认证

用户访问EMS服务控制台时，其本质是通过EMS服务管理面提供的REST风格的API接口进行请求。

EMS服务管理面的接口支持认证请求，需要用户从华为云统一身份认证服务 IAM获取正确的鉴权信息才能访问成功。关于IAM鉴权信息的详细介绍及获取方式，请参见**认证鉴权**。

访问控制

EMS默认资源隔离，IAM用户在EMS服务控制台创建的资源仅能被该IAM账号的管理员及其子用户访问。

6.3 数据保护技术

数据安全

EMS通过多种数据保护手段和特性，保障EMS数据安全可靠。

表 6-1 EMS 数据保护手段

数据保护手段	简要说明
传输加密（HTTPS）	为保证数据传输的安全性，访问EMS服务控制台时支持HTTPS协议。
操作认证	所有EMS服务管理面的API都会进行IAM身份认证。

审计与安全

出于分析或审计等目的，用户可以开启日志记录功能。通过将EMS日志规则配置到云日志服务 LTS，您可以获取EMS数据面的相关运行日志，从而帮助您定位问题。

服务韧性

EMS提供的是内存缓存，不是持久化存储，在EMS镜像重启/升级、节点重启、发生异常导致故障等场景下会导致内存缓存丢失，需要上层业务按缓存未命中进行处理。

监控安全风险

您可以通过在CCE集群上安装监控插件，将CCE集群的节点上部署的EMS监控数据上报至应用运维管理 AOM实例，便于您随时监控业务。可以通过配置监控告警规则，在出现EMS告警时，能够及时通知您处理告警。

故障恢复

EMS提供的是内存缓存，不是持久化存储，在EMS镜像重启/升级、节点重启、发生异常导致故障等场景下会导致内存缓存丢失，需要上层业务按缓存未命中进行处理。

由于EMS数据面采用融合部署，如果EMS数据面服务发生异常需要应急进行重启等运维操作或需要进行升级等变更操作时，需要用户授权或协助在用户容器集群上进行相关运维变更操作。

更新管理

由于EMS数据面采用融合部署，因此EMS数据面需要手动更新版本。如果EMS数据面服务需要进行升级更新操作时，需要用户授权或协助在用户容器集群上进行手动更新操作。

6.4 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行**申请下载**合规资质证书。

图 6-2 合规证书下载

资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看**资源中心**。

图 6-3 资源中心

销售许可证&软件著作权证书

另外，华为云还提供了以下销售许可证及软件著作权证书，供用户下载和参考。具体请查看**合规资质证书**。

图 6-4 销售许可证&软件著作权证书

7 计费说明

计费模式

EMS支持按需付费（后付费）计费方式。

按需付费（后付费）即先使用后付费的付费方式。您在华为云账户先充值，系统每小时统计前一小时的实际使用量并进行结算，从账户余额中扣除实际消费金额。

详细的服务资费费率标准请提工单咨询。

计费项

计费项为安装EMS业务集群节点使用的时长。

开始计费：EMS业务集群部署成功并激活后开始计费。

停止计费：EMS业务集群删除成功后停止计费。

计费方式

按节点实际使用的时长收费，以小时为单位，按每小时整点结算，不设最低消费标准。

计费公式

计费按照每小时计算，每小时内每个集群会按照00:05，00:10…，01:00时刻每隔5分钟采集在对应时间点上正常工作的节点数，称为打点数。

1个小时内的总打点数为12次采集的打点数之和。1小时内的总费用计算如下：

总费用=每小时每节点费用*Floor(总打点数/12)

说明

Floor()函数为向下取整函数。Floor(25/12)=2

计费示例一：

一个8节点的EMS集群从15:33开始运行，到16:00时对15:33-16:00 这段时间的费用开始计费。则一共有6个时间点（15:35,15:40,15:45,…,16:00），每个时间点有8个打点，所以总费用为Floor(6*8/12)*每小时每节点费用=4*每小时每节点费用

计费示例二：

一个8节点的EMS集群从15:08开始运行，到16:20的时候集群节点数扩展到16个节点，到17:21的时候集群节点数缩减到4个节点持续运行到17:41以后。则15:08-16:00,16:00-17:00,17:00-17:41三个小时各自的费用如下：

15:08-16:00：一共有11个时间点，每个时间点有8个节点，费用为Floor(11*8/12)*每小时每节点费用=7*每小时每节点费用

16:00-17:00：一共有12个时间点，16:05，16:10，16:15等3个时间点每个有8个节点，16:20到17:00等9个时间点有16个节点，费用为Floor((3*8+9*16)/12)*每小时每节点费用=14*每小时每节点费用

17:00-17:41：一共有8个时间点，17:05,17:10,17:15,17:20等4个时间点每个有16个节点，17:25,17:30,17:35,17:40等4个时间点每个有4个节点，费用为Floor((4*16+4*4)/12)*每小时每节点费用=6*每小时每节点费用

欠费

用户在使用EMS时，账户的可用额度小于待结算的账单，即被判定为账户欠费。欠费后，EMS会停止工作。已部署的EMS不会再产生作用。

8 约束与限制

EMS 权限管理

- IAM认证用户默认可以直接访问EMS控制台，无需授予任何IAM权限。
- 您需要使用通过EMS控制台申请的激活凭证激活EMS软件后才能正常使用EMS内存缓存功能。

EMS 使用须知

- EMS提供的是内存缓存，不是持久化存储，在EMS镜像重启/升级、节点重启、发生异常导致故障等场景下会导致内存缓存丢失，需要上层业务按缓存未命中进行处理。
- 为提高内存缓存性能，EMS内存缓存集群必须部署在同一AZ。
- EMS数据面镜像部署在用户的CCE容器集群上，EMS镜像运行需要占用AI节点的vCPU、内存等资源；同时EMS用于保存推理KVCache需要额外占用AI节点的内存资源。
- EMS数据面镜像部署在用户的CCE容器集群上，EMS镜像的日志、监控、告警需要对接云日志服务 LTS、应用运维管理 AOM等运维监控平台，并需要通过委托授权等方式将EMS镜像日志、监控等数据同步给EMS服务。

EMS 功能限制

- EMS提供的是内存缓存，不是持久化存储，在EMS镜像重启/升级、节点重启、发生异常导致故障等场景下会导致内存缓存丢失，需要上层业务按缓存未命中进行处理。
- 由于EMS采用融合部署，如果EMS数据面服务发生异常需要应急进行重启等运维操作、或需要进行升级等变更操作时，需要用户授权或协助在用户容器集群上进行相关运维变更操作。

9 与其他服务的关系

图 9-1 EMS 与其他服务的关系

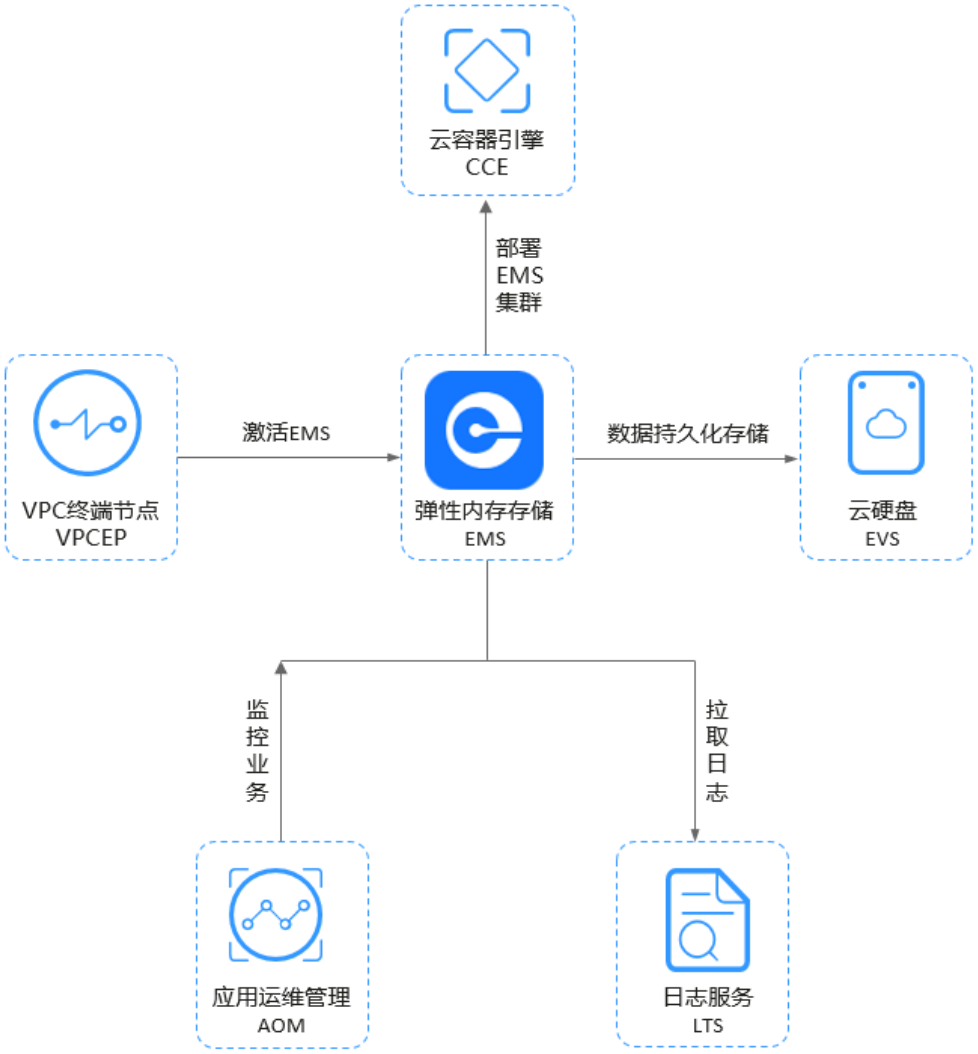


表 9-1 EMS 与其他服务的关系

服务名称	EMS服务与其他服务的关系	主要交互功能
云容器引擎（Cloud Container Engine，CCE）	EMS数据面镜像基于CCE云容器引擎服务进行安装部署。	部署EMS数据集群
云硬盘（Elastic Volume Service，EVS）	EMS使用云硬盘作为Zookeeper数据持久化存储。	部署EMS数据集群
VPC终端节点（VPC Endpoint，VPCEP）	通过VPC终端节点进行EMS激活和集群管理。	激活EMS
应用运维管理（Application Operations Management，AOM）	EMS将监控指标、告警等采集到应用运维管理平台，便于您随时监控业务。	收集运维指标
云日志服务（Log Tank Service，LTS）	EMS将运行日志转储到云日志服务，您可以获取EMS的相关操作日志，从而帮助您定位问题。	日志收集

10 基本概念

10.1 EMS 基本概念

KVCache

KVCache（Key-Value Cache）是用于加速大型语言模型（如Transformer模型）推理过程的技术，KVCache通过缓存Attention机制中的Key和Value矩阵（K和V），以避免在生成新token时重复计算历史序列的中间结果，减少冗余计算，从而显著提升推理效率。

LLM推理

LLM（Large Language Model）推理服务旨在为大规模语言模型（LLM）的推理任务提供高效、低延迟的在线服务能力。EMS通过KVCache缓存、多级缓存、分布式内存池化以及智能亲和调度等技术，加速推理速度并降低资源消耗。

激活凭证

您可以在EMS控制台创建激活凭证，您需要使用激活凭证激活EMS后才能正常使用EMS内存缓存功能。