弹性内存存储

常见问题

文档版本	01
发布日期	2025-06-19

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 概念类问题

1.1 什么是 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，从而显著提升了推理效率。

2 控制台使用类问题

2.1 为什么需要激活凭证

EMS采用半托管融合部署，EMS数据面部署在用户AI节点上，用户需要使用激活凭证激活EMS后才能开始使用。EMS激活时，EMS数据面会和EMS管理面通信，通过EMS管理面校验并完成用户关联后，用户才能使用EMS。

3 部署类问题

3.1 EMS 内存池需要占用 AI 节点多少 DRAM 内存

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。AI推理场景受限于显存瓶颈，DRAM内存富余较多，建议分配一半DRAM内存给EMS的KVCache内存池，EMS内存池空间越大，有利于提高KVCache缓存命中率，提升推理吞吐。

3.2 在执行主机配置脚本的过程中，无返回信息怎么办

问题现象

在执行主机配置脚本的过程中，当屏幕上显示如下信息时，可能会出现脚本一直无法结束的现象：

```
Modifying the kernel parameters.. If the process takes too long, restart the host to apply the changes.
```

可能原因

这通常是因为脚本在最后阶段会调整大页内核参数，将常规页面转换为大页。由于主机系统经过长时间运行后，物理内存可能出现严重的碎片化现象，内核需要整理并分配连续的大页内存区域。在分配大量大页内存的情况下，这一过程可能非常耗时。

解决方式

为了确保大页分配能够顺利进行并生效，建议可以考虑重启主机，利用系统初始化的过程来完成大页的重新分配与配置。

