弹性内存存储

常见问题

文档版本	01
发布日期	2026-05-25



华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2026。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

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
1.1 什么是 KVCache 1
2 控制台使用类问题 2
2.1 为什么需要激活凭证 2
3 部署类问题 3
3.1 EMS 内存池需要占用 AI 节点多少 DRAM 内存 3
3.2 在执行主机配置脚本的过程中，无返回信息怎么办 3
4 资源类问题 4
4.1 EMS 集群使用的安全组配置说明 4
4.2 配置 CoreDNS 插件 5
4.3 配置 LocalDNS 插件 6

 版权所有 © 华为云计算技术有限公司

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。

 版权所有 © 华为云计算技术有限公司

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。

版权所有 © 华为云计算技术有限公司

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。

 版权所有 © 华为云计算技术有限公司

4 资源类问题

4.1 EMS 集群使用的安全组配置说明

表 4-1 EMS 安全组推荐配置表

所属安全组	优先级	流量方向	协议类型	协议端口	源地址类型	源地址
ems-controller	1	入方向	TCP	8081	IP地址	VPC网段
ems-controller	1	入方向	TCP	5219	安全组	ems-server
ems-controller	1	入方向	TCP	5219	安全组	ems-controller
ems-controller	1	入方向	TCP	20051	安全组	默认容器子网安全组
ems-server	1	入方向	TCP	8081	IP地址	VPC网段
ems-server	1	入方向	TCP	8082	IP地址	VPC网段
ems-server	1	入方向	TCP	5220	安全组	ems-server
ems-server	1	入方向	TCP	5220	安全组	ems-controller
ems-server	1	入方向	TCP	20052	安全组	容器默认子网安全组

版权所有 © 华为云计算技术有限公司

所属安全组	优先级	流量方向	协议类型	协议端口	源地址类型	源地址
ems-server	1	入方向	TCP	5219	安全组	ems-server
ems-server	1	入方向	TCP	5219	安全组	ems-controller
ems-server	1	入方向	TCP	21859	安全组	ems-server
ems-server	1	入方向	TCP	21859	安全组	ems-controller
ems-server	1	入方向	TCP	20051	安全组	默认容器子网安全组
ems-server	1	入方向	TCP	21857	安全组	VPC网段
ems-zookeeper ems-zookeeper ems-zookeeper ems-zookeeper ems-zookeeper ems-zookeeper ems-zookeeper	1 1 1 1 1 1 1	入方向 入方向 入方向 入方向 入方向 入方向 入方向	TCP TCP TCP TCP TCP TCP TCP	10000 10000 11000 11000 10001 10002 20051	安全组 安全组 安全组 安全组 安全组 安全组 安全组	ems-server ems-controller ems-server ems-controller ems-zookeeper ems-zookeeper 默认容器子网安全组

4.2 配置 CoreDNS 插件

说明

DNS是K8s中至关重要的基础服务之一，当容器DNS策略配置不合理，集群规模较大时，DNS容易出现解析超时、解析失败等现象，极端场景下甚至会引起集群内业务大面积解析失败。

CoreDNS为Kubernetes集群提供DNS域名解析服务，在CCE集群创建时会默认安装，但需要根据集群规模进行CoreDNS的规格配置。

CoreDNS配置指导请见：**CoreDNS配置指导**。

EMS推荐的CoreDNS插件资源配额与CCE建议的CoreDNS插件资源配额一致，如**表4-2**所示：

版权所有 © 华为云计算技术有限公司

表 4-2 CoreDNS 插件配额建议

节点数量	推荐配置	实例数	CPU申请值	CPU限制值	内存申请值	内存限制值
50	2500QPS	2	500m	500m	512Mi	512Mi
200	5000QPS	2	1000m	1000m	1024Mi	1024Mi
1000	10000QPS	2	2000m	2000m	2048Mi	2048Mi
2000	20000QPS	4	2000m	2000m	2048Mi	2048Mi

4.3 配置 LocalDNS 插件

说明

随着集群规模增大，CoreDNS压力会增大，此场景下CoreDNS无法处理集群中的部分域名解析请求，导致该部分域名解析超时，影响集群的功能或者导致集群故障。LocalDNS通过在集群节点上作为守护程序集运行DNS缓存代理，提高集群DNS性能。

LocalDNS配置指导请见：**LocalDNS配置指导**。

版权所有 © 华为云计算技术有限公司