

PR #49966 完整报告

vllm-project/vllm

[Bugfix] Respect cgroup memory limits on all platforms

合并时间: 2026-07-28 14:38

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49966>

执行摘要

- 一句话: 修复 cgroup 内存限制在非 ROCm 平台被忽略
- 推荐动作: 值得快速合并。这是一个小而正确的 bugfix, 修复了容器场景下的常见问题。建议阅读时关注 `_get_available_ram_bytes` 和 `get_cgroup_memory_limit` 的配合逻辑。

功能与动机

用户报告在 Pod 中设置 memory limit 后, vLLM 仍然认为主机有 865 GB 可用内存, 从而错误地禁用 Auto-prefetch。根本原因是 `_get_available_ram_bytes` 仅在 ROCm 平台上读取 cgroup 限制, 其他平台直接返回 `psutil.virtual_memory().available`。

实现拆解

1. 在 `vllm/model_executor/model_loader/weight_utils.py` 中修改 `_get_available_ram_bytes` 函数: 移除 `if not current_platform.is_rocm(): return host_available` 的条件判断, 使所有平台都执行 cgroup 内存限制读取逻辑。
2. 更新函数 docstring, 从“honoring cgroup limits on ROCm”改为“honoring cgroup limits”, 反映通用化语义。
3. 该函数依赖 `get_cgroup_memory_limit`, 该函数已实现在 `vllm.utils.cpu_resource_utils` 中, 并支持跨平台 (Linux cgroup v1/v2), 因此无需新增依赖。

关键文件:

- `vllm/model_executor/model_loader/weight_utils.py` (模块 模型加载器; 类别 source; 类型 core-logic; 符号 `_get_available_ram_bytes`): 核心变更文件, 修改了 `_get_available_ram_bytes` 函数, 移除仅 ROCm 平台的 cgroup 限制检查, 使其在所有平台生效。

关键符号: `_get_available_ram_bytes`

关键源码片段

`vllm/model_executor/model_loader/weight_utils.py`

核心变更文件, 修改了 `_get_available_ram_bytes` 函数, 移除仅 ROCm 平台的 cgroup 限制检查, 使其在所有平台生效。

```
def _get_available_ram_bytes() -> int:
```

```

"""Return available RAM, honoring cgroup limits."""
import psutil

host_available = psutil.virtual_memory().available

# 从 cgroup 获取容器内存限制（适用于 Linux cgroup v1/v2）
from vllm.utils.cpu_resource_utils import get_cgroup_memory_limit

cgroup_limit, cgroup_usage = get_cgroup_memory_limit()
if cgroup_limit is None:
    # 非容器环境或无法读取 cgroup，直接返回主机可用内存
    return host_available
cgroup_available = (
    cgroup_limit if cgroup_usage is None else max(0, cgroup_limit - cgroup_usage)
)
# 取主机可用内存和容器可用内存的较小值
return min(host_available, cgroup_available)

```

评论区精华

审核人 AndreasKaratzas 评论“LGTM. Did not want to change the entire datapath, that's why I did not update it for other platforms in my PR.”表明原 ROCm PR 作者有意限制了平台范围，而当前 PR 将其通用化，得到认可。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅移除一条提前返回语句，逻辑在所有平台统一。
get_cgroup_memory_limit 已有对非 Linux 或非 cgroup 环境的 None 返回处理，不影响裸机运行。
- 影响：影响所有非 ROCm 的 Linux 容器用户（主要是 CUDA 用户），使其在 Pod 中设置的 memory limit 被正确读取，避免 Auto-prefetch 被错误禁用。裸机或非 Linux 平台无变化。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR