PR #40923 完整报告

vllm-project/vllm

[Kernel] Marlin MoE: include SM 12.x in default arch list

合并时间：2026-05-28 15:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40923

执行摘要

- 一句话：修复 SM12x 上 Marlin MoE 输出乱码
- 推荐动作：该 PR 值得精读，特别是 CUDA 架构标志管理和版本兼容性的模式（家族标志 vs 显式架构，以及版本门控）是 vLLM 构建系统中常见的实践。对于使用 Blackwell 硬件的团队，此修复是关键，建议尽快合入。

功能与动机

SM 12.x（如 RTX 50 系列、GB10/DGX Spark）上 Marlin-MoE 内核因缺乏原生 cubin，驱动使用 PTX JIT 回退，导致 MoE 输出乱码（V4-Flash *decode* 出现乱码 tokens），而 Hopper 上正常。PR 旨在将 SM12.x 加入编译架构列表以修复此 silent *correctness* bug。

实现拆解

1. 在 CMakeLists.txt 中修改四个 MARLIN 架构变量（MARLIN_ARCHS, MARLIN_BF16_ARCHS, MARLIN_FP8_ARCHS, MARLIN_MOE_ARCHS），添加 SM12.x 支持。
2. 对每个变量引入 CUDA 版本条件：若 CUDA >= 13.0 则使用家族标志 12.0f（单个 cubin 覆盖整个 SM12x 家族），否则显式列出 12.0a;12.1a（保证 CUDA 12.8 及以下兼容）。
3. 同步统一 MARLIN_FP8_ARCHS 的处理逻辑（之前硬编码为 8.9;12.0;12.1，现改为与其余变量相同的条件模式）。
4. 在 csrc/quantization/marlin/marlin.cu 中修改 W4A8-FP8 运行时检查，从精确判断 major_capability * 10 + minor_capability == 120 改为判断 major_capability == 12，使 SM121 等 SM12x 设备也能启用 FP8 Marlin 路径。
5. 添加独立的 CUDA 版本门控以保持与 CUDA 12.8 及以下兼容（参考 MLA_ARCHS 的现有模式）。

关键文件：

- CMakeLists.txt（模块 构建配置；类别 infra；类型 configuration）：核心变更文件，添加 SM12.x 到四个 Marlin 架构列表，根据 CUDA 版本选择家族标志或显式架构。
- csrc/quantization/marlin/marlin.cu（模块 量化内核；类别 source；类型 core-logic；符号 marlin_mm）：FP8 运行时检查放宽，以支持整个 SM12x 家族而不是仅 SM120。

关键符号：marlin_mm

关键源码片段

CMakeLists.txt

核心变更文件，添加 SM12.x 到四个 Marlin 架构列表，根据 CUDA 版本选择家族标志或显式架构。

```
# 在 CMakeLists.txt 中，每个 MARLIN 架构变量使用条件版本
# 家族标志 12.0f 仅 CUDA >= 13.0 支持，否则回退到 12.0a;12.1a

# marlin arches for fp16 output
if(${CMAKE_CUDA_COMPILER_VERSION} VERSION_GREATER_EQUAL 13.0)
  cuda_archs_loose_intersection(MARLIN_ARCHS "8.0+PTX;12.0f" "${CUDA_ARCHS}")
else()
  cuda_archs_loose_intersection(MARLIN_ARCHS "8.0+PTX;12.0a;12.1a" "${CUDA_ARCHS}")
endif()

# marlin arches for bf16 output
if(${CMAKE_CUDA_COMPILER_VERSION} VERSION_GREATER_EQUAL 13.0)
  cuda_archs_loose_intersection(MARLIN_BF16_ARCHS "8.0+PTX;9.0+PTX;12.0f" "${CUDA_
  ARCHS}")
else()
  cuda_archs_loose_intersection(MARLIN_BF16_ARCHS "8.0+PTX;9.0+PTX;12.0a;12.1a"
  "${CUDA_ARCHS}")
endif()

# marlin arches for fp8 input
if(${CMAKE_CUDA_COMPILER_VERSION} VERSION_GREATER_EQUAL 13.0)
  cuda_archs_loose_intersection(MARLIN_FP8_ARCHS "8.9;12.0f" "${CUDA_ARCHS}")
else()
  cuda_archs_loose_intersection(MARLIN_FP8_ARCHS "8.9;12.0a;12.1a" "${CUDA_ARCHS}")
endif()

# moe marlin arches
if(${CMAKE_CUDA_COMPILER_VERSION} VERSION_GREATER_EQUAL 13.0)
  cuda_archs_loose_intersection(MARLIN_MOE_ARCHS "8.0+PTX;12.0f" "${CUDA_ARCHS}")
else()
  cuda_archs_loose_intersection(MARLIN_MOE_ARCHS "8.0+PTX;12.0a;12.1a" "${CUDA_ARCHS}
  ")
endif()
```

csrc/quantization/marlin/marlin.cu

FP8 运行时检查放宽，以支持整个 SM12x 家族而不是仅 SM120。

```
// 在 marlin_mm 函数中，W4A8-FP8 路径的运行时检查
// 修改前：只允许 SM89 或 SM120
// 修改后：允许 SM89 或整个 SM12x 家族（major_capability == 12）
```

```
 if (a_type == vllm::kFE4M3fn) {
   TORCH_CHECK(major_capability * 10 + minor_capability >= 89,
               "FP8 only support Ada Lovelace or newer GPUs.");
   TORCH_CHECK(
       major_capability * 10 + minor_capability == 89 ||
       major_capability == 12,
       "Marlin W4A8-FP8 only support SM89 or SM12x device (It is slower than "
       "Marlin W4A16 on other devices).");
 }
```

评论区精华

- gemini-code-assist 建议同时添加 SM10.x，但 Harry-Chen 指出 Marlin 内核不用于 SM10.x（数据中心 Blackwell），未采纳。
- Harry-Chen 建议使用 12.0f 家族标志以减小 cubin 体积，但需注意该标志仅 CUDA >= 13.0 支持，最终添加了版本条件实现兼容性。
- Harry-Chen 提醒需放宽 marlin.cu 中的 FP8 运行时 SM120 检查，ubehera 提供了代码示例并采纳。
- 独立复现者 pasta-paul 和 idonati 分别在 RTX PRO 6000（SM120）和 DGX Spark 集群上验证了修复效果。
- 是否包含 SM10.x 架构 (design): 不添加 SM10.x，因为 Marlin 内核不用于该架构。
- 使用 12.0f 家族标志代替显式列表 (design): 采用条件性家族标志，兼顾 cubin 大小与旧 CUDA 兼容。
- 放宽 FP8 运行时 SM120 检查 (correctness): 运行时检查已修改为支持所有 SM12x 设备。

风险与影响

- 风险：主要风险是 CUDA 版本兼容性：12.0f 家族标志需 CUDA >= 13.0，已在 CMakeLists.txt 中添加版本门控，CUDA 12.8 及以下将回退到显式架构列表，无兼容问题。FP8 运行时检查放宽可能允许在 SM12x 上触发未充分验证的路径，但 Marlin 本身支持这些架构，仅性能可能不如原生优化，不会导致错误。对其他架构（如 sm_80/sm_90）无影响，因为 CMake 交叉编译自动过滤无关架构。
- 影响：影响范围：所有使用 SM12x 硬件（RTX 50 系列、GB10/DGX Spark）的用户，Marlin MoE 从输出乱码变为正确。对 Hopper 和其他架构无影响。构建系统改动仅涉及 CMakeLists.txt 和一处 CUDA 运行时检查，不会破坏现有构建。
- 风险标记：CUDA 版本条件依赖，仅验证 W4A16 路径

关联脉络

- PR #40760 [New Model] Support DeepseekV4: 关联 issue，DeepSeek V4 模型需要 SM12x 支持，本 PR 是其中的一个构建依赖。
- PR #40860 [Feat] DeepSeek V4 Rebased: 关联 issue，DeepSeek V4 的合入工作依赖 SM12x 内核编译支持。

- PR #40899 DeepSeek V4 support on SM12x with Triton sparse MLA fallback: 关联 issue，SM12x 上 DeepSeek V4 的另一种稀疏 MLA 回退方案，本 PR 修复的 Marlin 路径与其互补。