PR #35568 完整报告

vllm-project/vllm

[Bugfix] Fix SM121 (DGX Spark) exclusion from Marlin/CUTLASS FP8 paths

合并时间：2026-05-16 01:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/35568

执行摘要

- 一句话：修复 SM121 被排除在 Marlin/CUTLASS FP8 路径外
- 推荐动作：推荐阅读。该 PR 展示了如何通过有界家族匹配而非精确匹配来处理架构兼容性，是一种可复用的设计模式。同时解决了多个长期未关闭的 issue，对 Blackwell 用户至关重要。

功能与动机

SM121（DGX Spark）和 SM120（RTX 5090）具有相同的 FP8 MMA 能力，但被精确匹配的架构守卫排除，导致用户无法使用 Marlin/CUTLASS FP8 路径（issue #35432, #30163）。

实现拆解

1. 代码生成脚本（csrc/moe/marlin_moe_wna16/generate_kernels.py、csrc/quantization/marlin/generate_kernels.py）：将架构判断条件从 arch in [89, 120] 改为 arch == 89 or arch // 10 == 12，使 SM121 也能生成 FP8 内核模板。
2. 运行时 CUDA 检查（csrc/moe/marlin_moe_wna16/ops.cu）：将 TORCH_CHECK 中的精确 capability 比较改为检查 major_capability == 12。
3. CUTLASS dispatch 宏（csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm.cuh、scaled_mm_sm120_fp8_dispatch.cuh）：将 enable_sm120_only 替换为 enable_sm120_family，对应宏条件从 ==1200 改为 >=1200 && <1300。
4. Python 端输入验证（vllm/model_executor/layers/quantization/utils/marlin_utils.py）：将 is_device_capability(120) 替换为 is_device_capability_family(120)，并更新错误提示。
5. 测试文件（tests/kernels/moe/test_moe.py、tests/kernels/quantization/test_marlin_gemm.py）：使用 is_device_capability_family(120) 替换精确匹配；同时为 test_fused_marlin_moe 等三个测试添加 default_vllm_config fixture，修复因缺少配置上下文导致的失败。

关键文件：

- csrc/moe/marlin_moe_wna16/generate_kernels.py（模块 MOE 内核；类别 source；类型 core-logic；符号 SUPPORT_FP8）：核心代码生成脚本，控制 MOE Marlin 内核的 FP8 支持架构判断
- csrc/quantization/marlin/generate_kernels.py（模块 量化内核；类别 source；类型 core-logic；符号 SUPPORT_FP8）：与文件 1 类似，针对量化 Marlin 内核代码生成

- vllm/model_executor/layers/quantization/utils/marlin_utils.py（模块 量化工具；类别 source；类型 data-contract；符号 get_marlin_input_dtype）：Python 端 W4A8-FP8 输入类型检查，决定是否允许 FP8 路径
- tests/kernels/moe/test_moe.py（模块 MOE 测试；类别 test；类型 test-coverage；符号 marlin_moe_generate_valid_test_cases, test_fused_marlin_moe, test_fused_marlin_moe_with_bias, test_fused_marlin_moe_non_gated）：MOE 测试，更新架构判断以包括 SM12x 家族，并添加 missing vllm_config fixture
- tests/kernels/quantization/test_marlin_gemm.py（模块 量化测试；类别 test；类型 test-coverage）：量化 Marlin GEMM 测试，更新架构判断
- csrc/moe/marlin_moe_wna16/ops.cu（模块 MOE 内核；类别 other；类型 core-logic；符号 marlin_mm）：CUDA 运行时检查，决定是否允许 Marlin W4A8-FP8 执行
- csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm.cuh（模块 CUTLASS 调度；类别 other；类型 core-logic）：CUTLASS FP8 dispatch 主文件，使用宏控制 SM 范围
- csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm_sm120_fp8_dispatch.cuh（模块 FP8 调度；类别 other；类型 core-logic）：FP8 dispatch 实现，同样使用宏

关键符号：get_marlin_input_dtype, marlin_mm, marlin_moe_generate_valid_test_cases, test_fused_marlin_moe, test_fused_marlin_moe_with_bias, test_fused_marlin_moe_non_gated

关键源码片段

csrc/moe/marlin_moe_wna16/generate_kernels.py

核心代码生成脚本，控制 MOE Marlin 内核的 FP8 支持架构判断

```
# 从编译参数中解析架构列表
for arch in sys.argv[1].split(","):
    arch = arch[: arch.index(".") + 2].replace(".", "")
    arch = int(arch)
    # SM89 和 SM12x 系列 (SM120 RTX 5090, SM121 DGX Spark GB10)
    # 完全支持 mma.sync.aligned.m16n8k32.row.col.f32.e4m3.e4m3.f32
    # SM90 和 SM100 可通过 PTX 模拟，但无加速效果。
    # 原代码为 `if arch in [89, 120]`，现在使用有界家族匹配
    if arch == 89 or arch // 10 == 12:
        SUPPORT_FP8 = True
    if arch >= 80:
        SUPPORT_SM80 = True
    if arch == 75:
        SUPPORT_SM75 = True
```

评论区精华

主要讨论集中在 MOE 测试失败分析上：blake-snc 最初认为失败是预先存在的，后确认 fused_marlin_moe 在 set_current_vllm_config 上下文之外调用导致，最终通过添加 default_vllm_config fixture 修复。mgoin 询问失败原因，blake-snc 分析了日志并提交修复。

社区成员 AshtonVaughan 在 RTX 5090 上验证了家族检查逻辑，确认与 SM120 兼容。DavRodSwede 报告在 3 节点 DGX Spark 集群上运行 patched 镜像 38 天无问题。eugr 多次催促合并。

- MOE 测试失败分析及修复 (testing): 在三个测试函数上添加 @pytest.mark.usefixtures('default_vllm_config')
- SM12x 家族兼容性验证 (correctness): 逻辑验证通过，家族检查也覆盖 SM120
- 社区部署稳定性证据 (other): 补丁在真实生产环境中验证稳定

风险与影响

- 风险：风险较低。主要风险点：1）新的家族检查（arch // 10 == 12）可能在未来引入 SM13x 时意外匹配？但 PR 使用有界检查，不会匹配 SM13（130//10=13）。2）测试依赖真实硬件 SM121，CI 中无对应机型，覆盖率不足。3）添加 default_vllm_config fixture 可能影响测试独立性，但已与现有模式一致。
- 影响：对用户：DGX Spark（SM121）用户现在可以正常使用 Marlin FP4 和 CUTLASS FP8 路径，此前只能回退到慢速实现或报错。对系统：无性能回退，因为家族检查包含 SM120 且不引入额外开销。对团队：统一了 SM12x 架构处理方式，减少了未来添加新变体时的工作量。
- 风险标记：依赖真实硬件验证，架构家族边界风险，测试上下文依赖

关联脉络

- PR #30135 MXFP4 models still fall back to the Marlin kernel for RTX PRO 6000 (Blackwell SM120): 关联的 Blackwell 兼容性 issue，本 PR 扩展了 SM12x 家族支持
- PR #30163 Help Running NVFP4 model on 2x DGX Spark with vLLM + Ray (multi-node): 用户报告的 CUTLASS FP4 GEMM 失败问题，由本 PR 的 enable_sm120_family 变更解决
- PR #35432 Prebuilt vLLM wheels / official images fail on RTX 50-series (Blackwell, SM120/SM121): 此 PR 修复的 issue 之一，解决 SM121 上 FP8 模型运行时崩溃