

PR #32036 完整报告

sgl-project/sglang

[AMD] Minimax-M3 : unblock mxfp8 block convert on gfx950

合并时间: 2026-07-30 17:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32036>

执行摘要

- 一句话: 为 gfx950 添加 MXFP8 转 block-fp8 的 opt-in 开关
- 推荐动作: 该 PR 是一个典型的 opt-in 性能优化变更, 设计清晰、风险可控。建议合并后关注后续 cookbook 文档更新和 gfx950 原生 MXFP8 kernel 的进展。

功能与动机

在 gfx950 上, MXFP8 权重的 native MX dot_scaled 路径性能较差; 将其转换为 block-fp8 后, 可复用 gfx942 上的成熟内核, 实现 15-20% 的端到端吞吐提升, 且精度无损 (GSM8K 0.9719 vs 0.9689)。

实现拆解

1. 新增环境变量逻辑: 在文件 `python/sglang/srt/layers/quantization/fp8.py` 中, 修改模块级布尔变量 `_mxfp8_to_block_fp8_required` 的定义, 从只读 `mxfp8_block_convert_required()` 改为 `mxfp8_block_convert_required()` or `get_bool_env_var("SGLANG_FORCE_MXFP8_BLOCK_CONVERT")`。
2. 复用现有路径: 该变量控制后续权重量化初始化中是否触发 `convert_mxfp8_to_block` 转换, 以及是否使用 aiter 的 block-scale GEMM 和 fused MoE 内核。不引入新 kernel, 完全复用 gfx942 的代码路径。
3. 默认行为不变: 环境变量未设置时, `get_bool_env_var` 返回 `False`, `_mxfp8_to_block_fp8_required` 仅取决于原有的 `mxfp8_block_convert_required()` (`gfx942` 返回 `True`, `gfx950` 返回 `False`), 因此 `gfx950` 默认继续使用 native MX 路径。
4. 测试与文档: PR 未新增单元测试或文档, 但提供了详细的端到端 benchmark 和精度对比表。review 中建议后续补充 cookbook 文档。

关键文件:

- `python/sglang/srt/layers/quantization/fp8.py` (模块 量化层; 类别 source; 类型 core-logic): 核心变更文件, 新增环境变量开关以控制 MXFP8 到 block-fp8 的转换行为。

关键符号: 未识别

关键源码片段

`python/sglang/srt/layers/quantization/fp8.py`

核心变更文件，新增环境变量开关以控制 MXFP8 到 block-fp8 的转换行为。

```
# python/sglang/srt/layers/quantization/fp8.py
# gfx942 (MI300) has no MX matmul HW; MXFP8 checkpoints are converted to
# block-fp8 [128,128] at load and run through the native block-fp8 kernels.
# SGLANG_FORCE_MXFP8_BLOCK_CONVERT=1 opts into that same block-fp8 path on
# gfx950 (MI35x): it routes the fp8 GEMMs / fused MoE through the mature aiter
# block-scale kernels instead of the native MX dot_scaled path (measured +20%
# throughput at equal accuracy on MiniMax-M3, GSM8K 0.9719 vs 0.9689).
_mxfp8_to_block_fp8_required = mxfp8_block_convert_required() or get_bool_env_var(
    "SGLANG_FORCE_MXFP8_BLOCK_CONVERT"
)
```

评论区精华

审核人 HaiShaw 提出两条后续建议，未产生争议：

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅 7 行代码，属于模块级配置开关扩展，不修改任何运行时逻辑路径。环境变量未设置时行为与之前完全一致（byte-identical）。依赖现有 aiter 内核，无新增依赖。
- 影响：影响范围小，仅影响 gfx950 (MI35x) 上使用 MXFP8 量化（如 MiniMax-M3）的用户。需要手动设置环境变量才能激活新路径。CUDA 和 gfx942 路径完全不受影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #32230 [AMD] MiniMax-M3: opt-in custom/quick all-reduce on ROCm: 同一模型 MiniMax-M3 在 AMD 平台上的另一 opt-in 优化，共享相同的环境变量和性能调优模式。