PR #37018 完整报告

sgl-project/sglang

[Kernel] Fix SM90 FP8 decode regression with benchmarked M/K/N routing

合并时间：2026-08-30 07:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37018

执行摘要

- 一句话：限流 SM90 FP8 路由至大 M 包络，修复 decode 回归
- 推荐动作：值得精读。虽然仅 9 行改动，但它是"基准驱动路由选择"的教科书案例：先用宽网格确定优势面，再用随机化边界确认锁定阈值，最后对噪声区域采取保守策略。内核维护者和 FP8 性能工程师可以复刻这套验证协议；普通使用者了解结论即可——decode 一律走 AOT，大 prefill 宽投影走 Torch。建议后续为路由谓词补充形状注释表与轻量单测，防止回归。

功能与动机

PR body 明确指出 "The SM90 row/column-scaled FP8 selector introduced by #34318 can route decode GEMMs to torch._scaled_mm based on K/N alone"，带来 H100 W8A8 回归：模型 Llama-3.1-8B-Instruct-FP8-dynamic、batch size 1、吞吐约 215 → 约 190 tok/s，受影响下投影为 M=1, K=14336, N=4096。直接剖析显示该形状下 AOT 远快于 Torch（H200 上 26.50 µs 对 51.12 µs，AOT 快 1.93 倍；H100 上 32.19 µs 对 53.55 µs，AOT 快 1.66 倍），说明 decode 的 M 过小，Torch 后端的启动与张量化开销无法被大 K/N 掩盖，路由必须以 M 为先决条件。

实现拆解

1. 变更入口：唯一改动文件为 python/sglang/kernels/ops/gemm/__init__.py，修改 Fp8ScaledMMOp 依赖的 rowwise FP8 Torch/AOT 路由选择器 _prefer_torch_rowwise_fp8。
2. 谓词重构：旧谓词 (k >= 5376 and n >= 3584) or (k >= 3584 and m >= 8192) 的第一项完全不看 M，第二项也只是把 m >= 8192 与 k >= 3584 简单叠加；新谓词改为 m >= 8192 and ((k >= 4096 and n >= 6144) or (k >= 7168 and n >= 5376))，M 成为硬性前置门槛，K/N 包络按"宽输出投影"（QKV/gate/up 族）与"大 K down 投影"两类实测优势形状收窄。
3. 效果边界：M < 8192 的 decode 与小 prefill 全部留在 AOT；报告的 K=14336, N=4096 down 投影、窄 TP8 投影 5376x3584 / 3584x5376 以及未测量形状留在 AOT；M >= 8192 的宽 QKV/gate/up 与宽 down 投影继续走 Torch，保住大 prefill 收益。
4. 验证与配套：未新增测试文件，但 PR 给出了三阶段基准协议——132 案例宽网格（11 个投影族 × 12 个 M 值）、两次随机化边界确认（44/33 案例与 44 案例）、精确工作负载热身与 30~40 样本时序统计；H100 上新谓词选中的 32 个形状全部偏向 Torch（中位数

+19.48%），H200 同样全部偏向 Torch（中位数 +2.50%）。CI 侧重跑了受影响回归测试 test/registered/quant/test_w8a8_quantization.py，在 1-gpu-h100 上通过。

5. 提交演进：4 个 commit 显示从“Gate SM90 rowwise FP8 routing on M”到“Refine SM90 FP8 routing envelope”的两轮收敛，说明先加 M 门槛、再细化 K/N 包络的迭代过程，最后由 BBuf 修复 lint 并合入。

关键文件：

- python/sglang/kernels/ops/gemm/__init__.py（模块 路由选择；类别 source；类型 core-logic；符号 _prefer_torch_rowwise_fp8）：唯一改动文件，包含 Fp8ScaledMMOp 的 rowwise FP8 路由选择器 _prefer_torch_rowwise_fp8；本次修改的谓词直接决定 decode 与大 prefill GEMM 走 AOT 还是 torch._scaled_mm，是修复回归的核心。

关键符号：_prefer_torch_rowwise_fp8

关键源码片段

python/sglang/kernels/ops/gemm/__init__.py

唯一改动文件，包含 Fp8ScaledMMOp 的 rowwise FP8 路由选择器 _prefer_torch_rowwise_fp8；本次修改的谓词直接决定 decode 与大 prefill GEMM 走 AOT 还是 torch._scaled_mm，是修复回归的核心。

```
# python/sglang/kernels/ops/gemm/__init__.py
# Fp8ScaledMMOp 依赖此函数决定 rowwise FP8 GEMM 走 AOT 内核还是 Torch
def _prefer_torch_rowwise_fp8(m, k, n, ...):
    # 既有早期条件分支保持原样：不满足前置条件的形状在这里返回 False，
    # 由 AOT 内核执行，规避 Torch 后端在对应场景下的劣势

    # 本次修复的核心改动（替换了旧的 K/N-only 谓词）：
    # 1) m >= 8192 前置门槛：decode 与 M < 8192 的小 prefill 全部留在 AOT。
    # 实测 M=1、K=14336、N=4096 的 Llama-3.1-8B-Instruct-FP8-dynamic down
    # 投影在 H200 上 AOT 比 torch._scaled_mm 快 1.93 倍（26.50 μs 对 51.12 μs），
    # H100 上快 1.66 倍（32.19 μs 对 53.55 μs）；
    # 2) 只有在 M 足够大时，才把宽输出投影（gate/up/QKV，如 4096x6144、
    # 5376x21504）和宽 down 投影（如 14336x5376、28672x5376）路由到 Torch，
    # 这两类形状在 H100/H200 上均被 132 案例网格与随机边界确认证明更优；
    # 3) 窄投影（如 TP8 的 5376x3584、3584x5376）与未测量的形状留在 AOT，
    # 避免基于噪声调优造成二次回归。
    return (
        m >= 8192
        and ((k >= 4096 and n >= 6144) or (k >= 7168 and n >= 5376))
    )
```

评论区精华

本 PR 的讨论非常收敛：唯一正式 review 来自 BBuf（APPROVED）：“It seems resonable, approved!”，评审未提出技术质疑，改动以 PR 内大量基准数据自证。另一条线程是 CI 复验：通过 /rerun-test test/registered/quant/test_w8a8_quantization.py 在 1-gpu-h100 上通过

，但 pr-test-extra 与 AMD ROCm 网格状态为失败且 PR 内未解释原因。

- 路由改动合理性确认 (other)：已批准，无修改要求。
- W8A8 回归测试复验 (testing)：受回归影响的 W8A8 测试复验通过；extra/AMD CI 失败列入风险，未在合并前澄清。

风险与影响

- 风险：
  1. 路由谓词是未经单元测试锁定的魔数，后续重构 _prefer_torch_rowwise_fp8 可能无意破坏该包络而不被感知；建议补充覆盖“decode 必走 AOT、大 M 宽投影走 Torch”的轻量单测。
  2. 实测形状仅覆盖 11 个投影族 × 12 个 M 值，其他模型族或 TP 配置的未测量形状可能偏离最优。作者明确采用“不在噪声上路由”的保守策略，最坏情况是牺牲少量潜在收益，而非引入新的明显回归。
  3. m >= 8192 与 CUDA graph prefill 的 batch 上限耦合：未来若支持更大 prefill M，需要重做边界基准（M=8192 与 8320 的测量已显示收益在边界处规律变化）。
  4. pr-test-extra 与 AMD ROCm 网格状态失败且未在 PR 内澄清；虽然改动只影响 SM90 rowwise FP8 路由，理论上与 AMD 无关，但合并前未完全闭环。 - 影响：用户侧：H100 上 W8A8 模型小 batch 解码吞吐从约 190 tok/s 恢复到约 215 tok/s（约 12%）；大 prefill 用户仍享受 Torch 后端收益（H100 中位 +19.48%）。系统侧：影响范围仅限 rowwise FP8 GEMM 的后端选择，不触碰 AOT kernel 实现、非 FP8 路径或 CUDA graph 机制，属于低风险、高针对性的修复。团队侧：提供了一个可复用的“基准驱动路由调优”协议（宽网格 + 随机边界确认 + 精确工作负载热身），后续新架构或新形状的路由调整可以直接复用该验证方式。 - 风险标记：路由谓词无单元测试锁定，M 门槛依赖实测基准样本，未覆盖形状可能次优，CI extra/AMD 未闭环

关联脉络

- PR #34318 引入 SM90 FP8 路由选择器（上游 PR，标题未在材料中提供）：本 PR 的直接修复对象：#34318 引入的 K/N-only 路由导致 H100 decode 回归，PR body 引用了该 PR 下的 H100 W8A8 回归报告链接。