

PR #35297 完整报告

sgl-project/sglang

[BugFix][Qwen3.8] Support Qwen3.8-MXFP4 DCP by registering Qwen3_5 text-only archs in mamba radix cache whitelists

合并时间: 2026-08-25 13:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35297>

执行摘要

- 一句话: 修复 Qwen3.8-MXFP4 DCP 启动失败, 注册文本架构到 mamba 白名单
- 推荐动作: 该 PR 值得快速浏览, 它揭示了一个架构白名单与 DCP 配置交互的 bug。建议精读 `_MAMBA_EXTRA_BUFFER_ARCHS` 列表, 了解各架构的内存模型约束。未来应考虑为纯文本架构增加单元测试, 确保白名单同步。

功能与动机

Qwen3.8-2.4T-A95B 使用 Qwen3_5MoeForCausalLM 架构, 这是一个文本专用 Qwen3.5 MoE 架构, 但未在 `arg_groups/overrides.py` 的 mamba radix cache 白名单中列出。导致 `mamba_extra_buffer` 默认禁用, 当 DCP 启用时, 分配器页面大小按 DCP 大小缩放 (`page_size * dcp_size = 2`), 与 mamba 要求的 `page_size=1` 冲突, 启动报错 'AssertionError: MambaComponent requires page_size=1 when mamba_extra_buffer is disabled, got 2'。

实现拆解

1. 在 `python/sglang/srt/arg_groups/overrides.py` 中修改两个 `frozenset` 白名单:
 - `_MAMBA_RADIX_CACHE_ARCHS`: 在 `Qwen3_5ForConditionalGeneration` 后添加 `Qwen3_5MoeForCausalLM` 和 `Qwen3_5ForCausalLM`。
 - `_MAMBA_EXTRA_BUFFER_ARCHS`: 在 `Qwen3_5MoeForConditionalGeneration` 后添加同样的两个架构。
2. 动机是这些文本架构与 VL 变体共享相同的混合 GDN 主体, 且 `is_qwen3_5()` 已经包含全部四个架构, 因此应一致地访问 mamba radix cache。
3. 变更确保 `supports_mamba_cache_extra_buffer` 函数对这些架构返回 `True` (当 `linear_attn_backend` 为 `triton` 时), 从而使 `mamba_extra_buffer` 默认启用, 避免 DCP 场景下的 `page_size` 冲突。
4. 未新增测试文件, 但 PR 附带了 GSM8K 准确率测试结果 (准确率 0.931)。

关键文件:

- `python/sglang/srt/arg_groups/overrides.py` (模块 参数覆盖; 类别 `source`; 类型 `core-logic`; 符号 `_MAMBA_RADIX_CACHE_ARCHS`, `_MAMBA_EXTRA_BUFFER_ARCHS`) : 这是唯一变更文件, 核心逻辑是白名单注册, 直接影响 mamba radix cache 的启用和

DCP 启动。

关键符号: `supports_mamba_cache_extra_buffer`

关键源码片段

`python/sglang/srt/arg_groups/overrides.py`

这是唯一变更文件，核心逻辑是白名单注册，直接影响 mamba radix cache 的启用和 DCP 启动。

```
# python/sglang/srt/arg_groups/overrides.py
# 白名单集合: mamba radix cache 相关架构。
_MAMBA_RADIX_CACHE_ARCHS = frozenset(
    {
        "KimiLinearForCausalLM",
        "Qwen3_5ForConditionalGeneration",
        # 新增: 纯文本 Qwen3.5 MoE 架构, Qwen3.8-2.4T-A95B 使用此架构。
        "Qwen3_5MoeForCausalLM",
        "Qwen3_5ForCausalLM",
        # ... 其他架构
    }
)

# 白名单集合: 支持 extra_buffer mamba radix cache 策略的架构。
_MAMBA_EXTRA_BUFFER_ARCHS = frozenset(
    {
        "KimiLinearForCausalLM",
        "Qwen3_5ForConditionalGeneration",
        "Qwen3_5MoeForConditionalGeneration",
        # 新增: 使 mamba_extra_buffer 默认启用, 避免 DCP 下 page_size 冲突。
        "Qwen3_5MoeForCausalLM",
        "Qwen3_5ForCausalLM",
        # ... 其他架构
    }
)

# 检查模型架构是否支持 extra_buffer 策略 (纯读取)。
def supports_mamba_cache_extra_buffer(view: Any, model_arch: str) -> bool:
    # 若模型在 extra_buffer 白名单中, 且 linear_attn_backend 为 triton, 则返回 True。
    if model_arch in _MAMBA_EXTRA_BUFFER_ARCHS:
        return view.linear_attn_backend == "triton"
    return False
```

评论区精华

PR 只有 4 条评论，均为 CI 状态询问和 amd-bot 的自动回应。amd-bot 指出：本 PR 的代码未被任何 PR-CI 测试覆盖（唯一的覆盖测试是 nightly-only），各厂商流水线不完整，因此 CI 变绿不能证明修复有效。审查者 HaiShaw 批准了 PR。

- CI 覆盖不足 (testing): 不能仅依赖 CI 信号合入, 但 PR 最终仍被批准。

风险与影响

- 风险: 风险较低: 改动为 whitelist 添加两个字符串, 不会影响现有架构。潜在风险是已注册架构可能在 triton linear_attn_backend 下启用 mamba_extra_buffer, 但需要确认这些架构的实际后端支持, 否则可能引入新的启动错误。由于没有新增测试, 回归风险未被自动化验证。
- 影响: 影响范围限于启用 DCP 的 Qwen3.8-2.4T-A95B 部署, 使其能正常启动。对已有架构 (Kimi、Qwen3_5 VL 等) 无影响, 因为这些白名单是纯 additive。团队需注意 nightly 测试对本改动的覆盖。
- 风险标记: 白名单变更, 缺少测试覆盖, CI 未验证

关联脉络

- PR #36237 [MegaMoE] Respect padded MXFP8 scale row strides in pre-dispatch: 两者都涉及 Qwen3.5/MegaMoE 相关架构, 且都是对混合 MoE 的微小修复。