PR #31334 完整报告

sgl-project/sglang

[CPU] Fix mxfp4 padding size

合并时间：2026-07-21 09:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31334

执行摘要

- 一句话：修复 CPU 上 MXFP4 权重的 padding 尺寸
- 推荐动作：值得合并，修复明确且简洁。建议后续增加针对 MXFP4 + CPU AMX 的回归测试。

功能与动机

PR body 指出先前 PR #20072 缺少针对 MXFP4 的特定处理，导致 MoE 内核因 AMX 打包条件不满足而回退到原生路径，影响 GPT-OSS 120B/20B 模型性能。

实现拆解

1. 调整 intermediate padding size（python/sglang/srt/configs/update_config.py，+5 行）：在 adjust_config_with_unaligned_cpu_tp 函数中，计算 intermediate_padding_size 后增加 MXFP4 量化检测分支，将 padding size 翻倍。原因是 2 个 MXFP4 值打包为 1 个 uint8，需确保 intermediate size 对齐到 2 的倍数。
2. 移除原生回退路径（python/sglang/srt/layers/quantization/mxfp4.py，-23 行）：删除了 CPU AMX 条件下 use_intel_amx_backend 失败后的原生 dequantize 回退逻辑。该回退路径在 padding 正确后不再需要，且会绕过 AMX 加速。

关键文件：

- python/sglang/srt/configs/update_config.py（模块 配置；类别 source；类型 core-logic）：核心修复：在 intermediate padding size 计算中增加 MXFP4 量化检测，将 padding size 翻倍以确保 2 值 1 字节打包正确对齐。
- python/sglang/srt/layers/quantization/mxfp4.py（模块 量化；类别 source；类型 dependency-wiring）：移除不再需要的原生回退路径，该路径在 padding 正确后不再触发，删除可简化代码并防止误用。

关键符号：未识别

关键源码片段

python/sglang/srt/configs/update_config.py

核心修复：在 intermediate padding size 计算中增加 MXFP4 量化检测，将 padding size 翻倍以确保 2 值 1 字节打包正确对齐。

```
# python/sglang/srt/configs/update_config.py 第 254-259 行
intermediate_padding_size = tp_size * get_moe_padding_size(weight_block_size)
if model_config.quantization == "mxfp4":
    # For mxfp4 quantization, 2 mxfp4 values are packed to 1 uint8,
    # so we need to double the intermediate padding size to ensure
    # the padded intermediate size is divisible by 2 for proper packing.
    intermediate_padding_size *= 2
```

python/sglang/srt/layers/quantization/mxfp4.py

移除不再需要的原生回退路径，该路径在 padding 正确后不再触发，删除可简化代码并防止误用。

```
# python/sglang/srt/layers/quantization/mxfp4.py 第 837-883 行
elif _is_cpu and _is_cpu_amx_available:
    _amx_process_weight_after_loading(layer, ["w13_weight", "w2_weight"])
    if use_intel_amx_backend(layer):
        # ... 原始 AMX 打包代码 ...
        return
    # 以下回退路径已被删除:
    # from sglang.srt.layers.quantization.mxfp4_tensor import MXFP4QuantizeUtil
    # w13_weight = MXFP4QuantizeUtil.dequantize(...)
    # ...
    # return
else:
    # 非 CPU AMX 路径保持不变
    from triton_kernels.numerics_details.mxfp import upcast_from_mxfp
    ...
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险低。移除的回退路径仅在 AMX 打包失败时触发，而 padding 修正后 AMX 条件应始终满足。但若某些模型或配置下 AMX 仍不可用，则可能导致运行时崩溃（无备用路径）。建议至少在 CI 中覆盖 GPT-OSS 模型测试。
- 影响：影响范围仅限于 CPU 平台上使用 MXFP4 量化的 MoE 模型（如 GPT-OSS 120B/20B）。修复后可正确启用 AMX 打包，提升推理性能。对其他平台或量化格式无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #20072 [CPU] Add padding for mxfp4_weight: 本 PR 修复了该 PR 中遗漏的 MXFP4 特定 padding 处理。

- PR #29497 [CPU] Fix mxfp4 native fallback path: 本 PR 进一步回退了该 PR 中引入的回退路径，该路径在 padding 正确后不再需要。