

PR #35611 完整报告

sgl-project/sglang

[AMD] Enable moe_a2a_backend=mori for DeepSeek-V4 prefill context parallelism

合并时间: 2026-08-27 17:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35611>

执行摘要

- 一句话: 为 DeepSeek-V4 CP 启用 Mori A2A 后端
- 推荐动作: 该 PR 值得精读, 因为它展示了一个最小化、聚焦的后端启用模式。值得关注的设计决策包括: 将门控放宽在 hook 和模型侧两处同步修改, 以及在不改动核心逻辑的前提下信任 Mori 与 DeepEP 路径的一致性。若后续要扩展其他后端, 可参考此模式。

功能与动机

DeepSeek-V4 的预填充上下文并行此前会拒绝 `moe_a2a_backend=mori`, 尽管 Mori 与 DeepEP 使用相同的 rank-local EP 分发 / 合并路径。PR 描述中明确说明, 这两个 CP 门控排除 `mori` 的原因是“未经测试”而非“不兼容”。该变更旨在让用户能够在 CP 场景下使用 Mori, 从而充分利用其性能优势。

实现拆解

实现拆解

1. 扩展 CP 校验白名单: 在 `python/sglang/srt/arg_groups/deepseek_v4_hook.py` 的 `validate_deepseek_v4_cp` 中, 将 `mori` 添加到 `supported_a2a_backends` 元组中, 并更新错误消息以反映新的支持集合。
2. 放宽模型侧后端门控: 在 `python/sglang/srt/models/deepseek_v4.py` 的 `_run_moe_ffn_dp_sync` 中, 将 `mori` 加入 `is_deepep()` or `is_megamoe()` 的断言, 使 CP 分支允许 `mori` 后端。
3. 保留原有操作语义: 不修改调度器、集合通信或计算内核。Mori 路径仍走 `forward_deeppep` 的 `combine` 逻辑, 不执行 `a2a=None` 时的 CP gather/reduce-scatter, 因此复制的共享专家只添加一次。
4. 配置与测试配套: 无直接测试文件变更; 通过 8xMI355X 环境的手工验证 (40/40 稳定首 token 匹配, 4K-16K 长度匹配) 来保障正确性。

关键文件:

- `python/sglang/srt/models/deepseek_v4.py` (模块 模型运行; 类别 source; 类型 data-contract; 符号 `_run_moe_ffn_dp_sync`): 模型侧核心路径, 修改 CP 分支的 A2A 后端断言, 允许 `mori`, 是使 CP 与 Mori 兼容的关键。

- python/sglang/srt/arg_groups/deepseek_v4_hook.py (模块 参数校验; 类别 source; 类型 core-logic; 符号 validate_deepseek_v4_cp) : 参数校验 hook, 更新 CP 的白名单以包含 mori, 确保用户配置能通过校验。

关键符号: _run_moe_ffn_dp_sync, validate_deepseek_v4_cp

关键源码片段

python/sglang/srt/models/deepseek_v4.py

模型侧核心路径, 修改 CP 分支的 A2A 后端断言, 允许 mori, 是使 CP 与 Mori 兼容的关键。

```
# python/sglang/srt/models/deepseek_v4.py 中的 CP 分支
if _use_cp:
    moe_a2a_backend = get_moe_a2a_backend()
    if moe_a2a_backend.is_none():
        hidden_states = dsa_cp_gather_hidden_states(hidden_states)
    else:
        # 允许 deepseek、megamoe 和 mori 作为 CP 的 A2A 后端
        assert (
            moe_a2a_backend.is_deepseek()
            or moe_a2a_backend.is_megamoe()
            or moe_a2a_backend.is_mori()
        ), (
            "CP requires moe_a2a_backend in ('deepseek', 'megamoe', 'mori'), "
            f"got {moe_a2a_backend.value!r}."
        )
```

评论区精华

Review 评论为空, 但 PR 作者在 body 中详细说明了安全性论证。审查者 1am9trash 批准了 PR, 并评价“Only release gate to make mori runnable with pcp.”, 确认这是一次最小化的门控放开。

- 暂无高价值评论线程

风险与影响

- 风险: 技术风险:
 - 代码路径风险: CP + Mori 组合此前未在代码中启用, 虽然 Mori 与 DeepEP 路径相似, 但在极端 token 分布下可能存在未预期的数值差异。
 - 共享专家计数风险: PR 声明 Mori 不受 #35535 影响, 因为其走 forward_deepseek_combine 后只添加一次共享专家; 但若未来对 a2a=None 路径的修复发生回归, 需警惕。
 - 性能风险: Mori 动态容量限制及 CP-v1 的 one-token-per-rank GPU 错误是已知外部问题, 可能影响性能或稳定性。
 - 测试覆盖缺失: 未添加自动化测试, 依赖手工验证, 存在回归风险。
 - 影响: 影响范围: 影响所有使用 DeepSeek-V4 预填充 CP 且希望使用 Mori 后端的用户。本变更未改变默认行为 (默认仍是 None), 因此对现有用户影响较小, 但对 AMD 平台

用户提供了新选择。对团队而言，需要维护新增的门控逻辑，并关注未来 CP+Mori 组合的稳定性。

- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #35535 Fix replicated (TP1) shared expert being summed by collectives that skip the MoE all-reduce: 本 PR 的安全论证依赖 #35535 中关于 Mori 不受 CP gather/reduce-scatter 影响的分析，两者在共享专家处理上相关。