PR #34715 完整报告

sgl-project/sglang

[bugfix] [NPU] fix transpose batch matmul K*B exceed 65536.

合并时间：2026-08-24 15:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34715

执行摘要

- 一句话：修复 NPU 转置批量矩阵乘超限崩溃，迁移 K3 配置
- 推荐动作：值得精读。重点关注 forward_mla_core_npu 的回退分支设计以及 review 中如何把 K3 专属参数泛化为通用 ServerArgs 参数。建议后续补充针对大形状边界的单元测试或 NPU 回归测试。

功能与动机

PR body 指出：torch_npu.npu_transpose_batchmatmul 提供了 Kimi-K3 数值验证过的路径，但不支持相关维度达到限制的形状，特别是 B * K 必须小于 65536；大 prefill 工作负载会超出该限制并在运行时失败。另外，为了让配置入口更统一，PR 将 SGLANG_K3_SHARED_EXPERTS_ATTN_TP 与 SGLANG_K3_DENSE_MLP_ATTN_TP 两个环境变量迁移为 ServerArgs 选项。

实现拆解

1. NPU MLA 核心路径加入回退分支（python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py）：在 forward_mla_core_npu 中，attn_output.view(-1, num_local_heads, kv_lora_rank) 之后、o_proj 之前，增加三个维度上限判断：attn_output.shape[0]（batch 维）、attn_output.shape[-1] * attn_output.shape[-2]（K 维乘积）、m.w_vc.shape[-1]（权重 K 维），任一超过 65536 时预分配输出缓冲并调用 torch.ops.npu.batch_matmul_transpose；否则维持 torch_npu.npu_transpose_batchmatmul 调用。这样既保留正常形状的数值验证路径，又解除大 prefill 的运行时崩溃。
2. 配置入口收口（python/sglang/srt/environ.py）：删除 SGLANG_K3_SHARED_EXPERTS_ATTN_TP 与 SGLANG_K3_DENSE_MLP_ATTN_TP 两个 EnvBool 定义，避免环境变量与 ServerArgs 双入口并存导致的行为漂移。
3. 新增通用 ServerArgs 参数（python/sglang/srt/server_args.py）：在 ServerArgs 的 parallel 命名空间下新增 enable_shared_experts_attn_tp 与 enable_dense_mlp_attn_tp 两个布尔参数，帮助信息分别说明共享专家权重与 dense MLP 权重在 attention-TP 组上的切分语义。命名最初带 k3_ 前缀，经 review 泛化为 enable_*，便于其他模型复用。
4. Kimi-K3 模型内改用 ServerArgs 读取（python/sglang/srt/models/kimi_k3.py）：移除模块级 _k3_shared_experts_attn_tp、_k3_dense_mlp_attn_tp 读取，改为运行时通过 get_parallel().enable_shared_experts_attn_tp 与

get_parallel().enable_dense_mlp_attn_tp 读取。KimiK3MLP.__init__中的_dense_attn_tp、_shared_experts_tp1 与 _shared_experts_attn_tp_comm 三个分支判定均改为读取新字段，逻辑保持不变。

5. 配套验证：PR body 报告 Kimi-K3 gsm8k 准确率 98%（200 题）；未新增单元测试文件，checklist 未勾选测试项，快速验证依赖 NPU CI。

关键文件：

- python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py（模块 NPU 内核；类别 source；类型 core-logic；符号 forward_mla_core_npu）：NPU MLA 注意力核心前向入口，本 PR 在此加入 65536 上限判断与回退逻辑，是修复的主体。
- python/sglang/srt/models/kimi_k3.py（模块 K3 模型；类别 source；类型 data-contract；符号 KimiK3MLP.init）：Kimi-K3 模型实现，将环境变量读取改为 ServerArgs，驱动共享专家与 dense MLP 的 attention-TP 布局。
- python/sglang/srt/server_args.py（模块 服务参数；类别 source；类型 core-logic；符号 ServerArgs.enable_shared_experts_attn_tp, ServerArgs.enable_dense_mlp_attn_tp）：新增通用 attention-TP 参数，是配置迁移的核心，命名经 review 泛化。
- python/sglang/srt/environ.py（模块 环境配置；类别 source；类型 core-logic；符号 Envs.SGLANG_K3_SHARED_EXPERTS_ATTN_TP, Envs.SGLANG_K3_DENSE_MLP_ATTN_TP）：删除 K3 专属环境变量定义，完成配置入口收口。

关键符号：forward_mla_core_npu, KimiK3MLP.init

关键源码片段

python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py

NPU MLA 注意力核心前向入口，本 PR 在此加入 65536 上限判断与回退逻辑，是修复的主体。

```
def forward_mla_core_npu(
    m: 'DeepseekV2AttentionMLA',
    q_pe: torch.Tensor,
    k_pe: torch.Tensor,
    q_nope_out: torch.Tensor,
    k_nope: torch.Tensor,
    forward_batch: 'ForwardBatch',
    zero_allocator: 'BumpAllocator',
    positions: torch.Tensor,
    topk_indices: torch.Tensor,
) -> torch.Tensor:
    attn_output = m.attn_mqa(
        q_nope_out,
        k_nope,
        k_nope,
        forward_batch,
        q_rope=q_pe,
```

```
        k_rope=k_pe,
        **(dict(topk_indices=topk_indices) if topk_indices is not None else {}),
    )

    attn_output = attn_output.view(-1, m.num_local_heads, m.kv_lora_rank)
    attn_output = attn_output.contiguous()

    # torch_npu.npu_transpose_batchmatmul 是 Kimi-K3 数值验证过的路径，
    # 但其内部限制任意维度必须小于 65536；大 prefill 会触发运行时失败。
    # 这里对 batch 维、K 维乘积以及权重 K 维逐一做上限检查。
    if (
        attn_output.shape[0] >= 65536
        or attn_output.shape[-1] * attn_output.shape[-2] >= 65536
        or m.w_vc.shape[-1] >= 65536
    ):
        # 超出上限时回退到 torch.ops.npu.batch_matmul_transpose，
        # 先分配输出缓冲，再调用带 out 参数的算子，保证大形状可用。
        attn_bmm_output = torch.empty(
            (attn_output.shape[0], m.num_local_heads, m.v_head_dim),
            dtype=attn_output.dtype,
            device=attn_output.device,
        )
        torch.ops.npu.batch_matmul_transpose(attn_output, m.w_vc, attn_bmm_output)
    else:
        # 常规大小继续走数值验证过的实现，保持 Kimi-K3 正常路径的精度。
        attn_bmm_output = torch_npu.npu_transpose_batchmatmul(
            attn_output,
            m.w_vc,
            perm_x1=(1, 0, 2),
            perm_x2=(0, 1, 2),
            perm_y=(1, 0, 2),
        )

    attn_bmm_output = attn_bmm_output.reshape(-1, m.num_local_heads * m.v_head_dim)
    output, _ = m.o_proj(attn_bmm_output)
    return output
```

python/sglang/srt/server_args.py

新增通用 attention-TP 参数，是配置迁移的核心，命名经 review 泛化。

```
# 位于 ServerArgs 的 parallel 命名空间下，供所有使用 expert-parallel
# all-to-all 后端的模型复用，不再限定 K3。
enable_shared_experts_attn_tp: A[
    bool,
    'Shard shared expert weights across the attention TP group when using an expert-parallel all-to-all backend.',
    NS('parallel'),
] = False

# 在 DP attention 下，将 dense MLP 权重按 attention TP 组切分，
```

```
# 允许 NPU 启动器保留已验证的 attention-TP 布局。
enable_dense_mlp_attn_tp: A[
    bool,
    'Shard dense MLP weights across the attention TP group under DP attention.',
    NS('parallel'),
] = False
```

评论区精华

Reviewer Hexq0210 在 server_args.py review 中连续提出命名与设计建议：Do not add a separate parameter for K3. Instead, write a shared parameter name that can be reused by other models. 随后又给出具体命名 enable_shared_experts_attn_tp 与 enable_dense_mlp_attn_tp。作者通过 generalize attention TP server args 与 clarify attention TP sharding options 两次提交采纳建议，将 K3 专属参数泛化为通用参数，这是本 PR 最具价值的 review 决策。

- K3 专属参数应泛化为通用参数 (design): 作者通过 generalize attention TP server args 与 clarify attention TP sharding options 两次提交采纳，最终命名为 enable_shared_experts_attn_tp / enable_dense_mlp_attn_tp。
- 参数命名建议 (design): 最终 server_args.py 中的字段名即为这两个名称。

风险与影响

- 风险：
  - 数值一致性风险：torch.ops.npu.batch_matmul_transpose 在 base 注释中曾被标记为对 Kimi-K3 数值不等价，现在大形状会切换到该路径，精度差异需要关注；PR 仅用 200 题 gsm8k 验证，覆盖有限。
  - 边界判断覆盖风险：代码用三个维度的 >= 65536 判断近似 B * K 限制。如果上游张量的 batch 维与 K 维各自小于 65536 但乘积超过，可能漏判；需要结合算子文档确认。
  - 配置破坏性变更：删除两个环境变量会影响依赖旧变量启动的脚本，且最终 CLI 参数名与 PR body 中声明的 --k3-不同（实际为 --enable-），文档与部署脚本需同步更新。
  - 测试缺口：改动无对应单元测试，NPU 路径的回归主要依赖硬件 CI，跨平台风险相对可控（仅 NPU 分支生效）。
- 影响：
  - 用户影响：NPU 后端跑 Kimi-K3 大 prefill 的用户不再因 batch matmul 形状超限崩溃，但需按新参数名调整启动脚本。
  - 系统影响：改动仅作用于 NPU 分支的 MLA 注意力路径，GPU/CPU 等其他后端不受影响。
  - 团队影响：配置入口从环境变量迁移到 ServerArgs，后续其他模型可复用 attention-TP 布局选项，减少 NPU 专属分支。
  - 影响程度：中等偏低，限定在 NPU 与 Kimi-K3/DeepSeek MLA 相关场景。
  - 风险标记：NPU 核心路径变更，缺少测试覆盖，配置接口破坏性变更，边界判断覆盖不全

关联脉络

- PR #35508 [NPU] [DOC] Add Ascend NPU (A3) recipe to the Kimi-K3 cookbook: 同为 Kimi-K3 在 Ascend NPU 上的支持工作，文档 recipe 与本 PR 的 NPU 注意力修复互补。
- PR #36124 [AMD] Quark shared-experts gate: recognise a trailing MTP layer: 均涉及共享专家（shared experts）的 TP 布局判定，平台不同但配置语义相近。