PR #27053 完整报告

sgl-project/sglang

[BCG][GLM5] perf: BCG support and prefill enhancements

合并时间：2026-06-25 04:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27053

执行摘要

- 一句话：为 GLM5 引入 BCG 支持并优化 prefill 性能
- 推荐动作：强烈建议精读。本 PR 展示了如何在推理框架中平衡 *eager* 与 CUDA 图执行，是 CUDA 图后端演进的关键步骤。特别关注 MlaBmmFusionPlan 和 mla_bmm_then_unified_attention 的设计，以及在 split op 中灵活切换 *eager* 与图的策略。讨论中的设计取舍值得团队参考。

功能与动机

原 PR #23351 为 GLM5（nsa_indexer）引入 PCG，但由于索引器完整路径不能完全被 CUDA 图捕获，导致 *eager* 模式实际性能更优。且 PCG 拆分产生单 kernel 图岛，带来不必要的调度开销。为提升 prefill 性能，本 PR 提出 BCG 支持及相关优化。

实现拆解

1. 引入 BCG 上下文感知：在 breakable_cuda_graph 模块中添加 call_with_graph_break 函数，允许在图中标记 *eager* 执行点。
2. 重构 DSA 索引器：将 k_cache_and_topk_result 拆分为 dsa_indexer_graph_dispatch（PCG/BCG 共用）和 bcg_k_cache_and_topk_result（BCG 特有），并将 logits_head_gate 注册为 split op，在图捕获时 *eager* 执行。
3. 新增 MLA BMM 融合计划：定义 MlaBmmFusionPlan dataclass，实现 _can_fuse_bmm_into_attention 判断条件，在满足条件时通过 mla_bmm_then_unified_attention 将 BMM 和 unified_attention 合并为一个 *eager* split op。
4. 支持 DeepSeek-V2 双流 MoE 图：在 deepseek_v2.py 中添加 _can_dual_stream_graph 逻辑，当条件满足时调用 dsv2_flashinfer_moe_dual_stream_graph 进行双流图捕获，通过环境变量 SGLANG_ENABLE_PCG_DSV2_DUAL_STREAM 控制。
5. 移除空 CUDA 图：在 BCG 段结束时增加 drop_empty 参数，若段内无 kernel 则不创建空图，减少警告。
6. 测试与验证：新增端到端测试 test_pcg_glm5_fp8_tp8.py（测试 BCG），在 8 卡 H200 上运行 GSM8K 评估，确保精度达标（>0.92）。

关键文件：

- python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py（模块 MLA 前向；类别 source；类型 data-contract；符号 MlaBmmFusionPlan, _can_fuse_bmm_into_attention, _split_q_nope_pe, _make_mla_bmm_fusion_plan）：定义了核心数据契约 MlaBmmFusionPlan，实现 BMM 与 unified_attention 融合的判断与执行，是 prefill 优化的关键路径。
- python/sglang/srt/layers/attention/dsa/dsa_indexer.py（模块 DSA 索引器；类别 source；类型 dependency-wiring；符号 _is_in_piecewise_or_breakable_cuda_graph, k_cache_and_topk_result, _logits_head_gate_pcg_fake_impl, _logits_head_gate_graph_fake_impl）：DSA 索引器图调度重构，引入统一的 dsa_indexer_graph_dispatch 作为 split op，并支持 BCG 的 eager_on_graph 变体。
- python/sglang/srt/models/deepseek_v2.py（模块 DeepSeek 模型；类别 source；类型 data-contract；符号 _can_dual_stream_graph, dsv2_flashinfer_moe_dual_stream_graph）：添加 dual-stream MoE 图捕获条件，支持 DeepSeek-V2 在 PCG/BCG 下双流异步执行。
- test/registered/cuda_graph/piecewise/test_pcg_glm5_fp8_tp8.py（模块 BCG 测试；类别 test；类型 test-coverage；符号 TestBCGGlm5Fp8TP8, setUpClass, tearDownClass, test_gsm8k）：新增端到端 BCG 测试，在 8 卡 H200 上覆盖 GLM5 FP8 模型，验证精度与性能。
- python/sglang/srt/layers/attention/dsa/utils.py（模块 DSA 工具；类别 source；类型 core-logic；符号 is_graph_dsa_split_op_surface）：新增 is_graph_dsa_split_op_surface 函数，统一 PCG/BCG 下 split-op 表面的判断逻辑。
- python/sglang/srt/layers/attention/dsa_backend.py（模块 注意力后端；类别 source；类型 dependency-wiring）：在预填充实现中增加 BCG 上下文判断，强制关闭 MHA 以确保图重放正确性。

关键符号：MlaBmmFusionPlan, _can_fuse_bmm_into_attention, _split_q_nope_pe, _make_mla_bmm_fusion_plan, mla_bmm_then_unified_attention, _is_in_piecewise_or_breakable_cuda_graph, k_cache_and_topk_result, dsa_indexer_graph_dispatch, logits_head_gate_pcg, logits_head_gate_graph, _should_skip_logits_computation, _can_dual_stream_graph, dsv2_flashinfer_moe_dual_stream_graph, is_graph_dsa_split_op_surface

关键源码片段

python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py

定义了核心数据契约 MlaBmmFusionPlan，实现 BMM 与 unified_attention 融合的判断与执行，是 prefill 优化的关键路径。

```
@dataclass(frozen=True)
class MlaBmmFusionPlan:
    """BMM 与 unified_attention 融合所需的所有预分配缓冲区。"""
    q_nope_t: torch.Tensor # 转置的 q_nope
    q_nope_out_buf: torch.Tensor # BMM 输出缓冲区（attn 输入）
```

```
    q_nope_out_view: torch.Tensor # 视图避免额外拷贝
    attn_output_buf: torch.Tensor # attention 输出缓冲区


def _can_fuse_bmm_into_attention(
    self: DeepseekV2AttentionMLA, forward_batch: ForwardBatch
) -> bool:
    # 仅在 graph DSA split-op 表面（非 spec 的 extend + 图模式）且满足条件下启用
    if not is_graph_dsa_split_op_surface(forward_batch):
        return False
    if not self.use_dsa:
        return False
    if self.use_deep_gemm_bmm or _is_hip:
        return False
    if is_kv_b_lora_active(self):
        return False
    # fp8 和 DeepGEMM 已经有各自融合路径，这里仅支持 bf16 回退
    if self.w_kc.dtype == torch.float8_e4m3fn:
        return False
    return True
```

python/sglang/srt/layers/attention/dsa/dsa_indexer.py

DSA 索引器图调度重构，引入统一的 dsa_indexer_graph_dispatch 作为 split op，并支持 BCG 的 eager_on_graph 变体。

```
def _is_in_piecewise_or_breakable_cuda_graph() -> bool:
    """判断当前是否在 PCG 或 BCG 图捕获中（用于 DSA 索引器 dispatch）。"""
    return is_in_tc_piecewise_cuda_graph() or is_in_breakable_cuda_graph()


GRAPH_WEIGHTS_PROJ_LORA_ERROR = (
    "DSA indexer weights_proj LoRA is incompatible with "
    "piecewise/breakable CUDA graph; remove the explicit "
    "prefill cuda-graph backend override or drop "
    "indexer.weights_proj from the LoRA target modules."
)


# 原单一 k_cache_and_topk_result 被重构为两个变体：PCG 保留原有裁剪逻辑，
# BCG 通过 eager_on_graph 在 BCG 中标记为 eager 执行，避免空图。
@register_custom_op(mutates_args=["topk_result"])
@register_split_op()
def k_cache_and_topk_result(
    layer_id: int, key: torch.Tensor, q_fp8: torch.Tensor,
    weights: torch.Tensor, topk_result: torch.Tensor,
) -> None:
    # 具体实现在 PR 中展开
    pass
```

```
# BCG variant: 用 eager_on_graph 包装使该 op 在图捕获时始终 eager 执行
bcg_k_cache_and_topk_result = eager_on_graph(True)(k_cache_and_topk_result)
```

评论区精华

关键讨论聚焦于代码复杂度与设计取舍：

- 形状不匹配修复：gemini-code-assist[bot] 指出 dsa_indexer_graph_dispatch 中 q_fp8[:extend_num_tokens] 与未切片的 weights 形状不匹配。作者在后续 commit 中统一切片方式修复。
- BCG 拆分之争：Oasis-Git 认为 BCG 部分可移到独立 PR 以降低复杂度，作者回应已简化设计并移除冗余 env flag，最终获得批准。
- 代码组织：Oasis-Git 建议 PCG/BCG 函数应放在文件底部而非 _is_cuda 块内，作者采纳。
- 空 CUDA 图收益：Oasis-Git 询问 drop_empty 是否必要，作者解释主要收益来自 single-kernel 图合并，移除空图是边际优化但可消除警告。
  - dsa_indexer_graph_dispatch 中 q_fp8/weights 形状不匹配 (correctness): 作者在后续 commit 中统一切片方式，确保 weights 也切片至相同长度。
  - BCG 代码复杂度及拆分为独立 PR 的讨论 (design): 当前 PR 保留 BCG 支持且合并，但未来可能进一步分离。
  - PCG/BCG 相关函数的代码组织风格 (style): 作者采纳建议，将函数移至文件底部。
  - 空 CUDA 图处理的性能收益 (performance): 功能保留且未被推翻。

风险与影响

- 风险：
  1. 回归风险：BCG 与 PCG 路径并存，部分配置可能未充分测试；除 GLM5 FP8 外，其他 DSA 模型（如 DeepSeek-V3、Janus）未覆盖。
  2. 性能风险：split op 引入 eager 执行可能增加 host-device 同步点，在长序列场景下可能抵消部分收益。
  3. 兼容性风险：SGLANG_ENABLE_PCG_DSV2_DUAL_STREAM 环境变量控制的双流 MoE 图与部分 MoE 后端（A2A、EPLB）不兼容，已在条件中排除。
  4. 维护风险：PR 含 83 次提交，代码复杂度较高，后续维护成本较大。
- 影响：
  1. 用户：GLM5 用户直接获得约 4% 吞吐提升和 10% TTFT 降低；DeepSeek-V2 用户可设置环境变量启用双流 MoE 图降延迟。
  2. 系统：BCG 后端成为 prefill 图捕获的新选项，计划逐步取代 PCG，需要迁移。
  3. 团队：需管理 PCG、BCG 两套图后端，维护成本增加；同时需关注与 DSA 索引器和 MLA 融合路径的兼容性。 - 风险标记：核心路径变更，硬件覆盖有限，环境变量依赖，空图警告噪音

关联脉络

- PR #23351 [PCG] Add piecewise CUDA graph support for NSA (GLM5) models: 本 PR 是在 #23351 引入的 PCG 基础上进行的 BCG 优化和性能增强，PR body 中明确引用 #23351 作为动机。