PR #25885 完整报告

sgl-project/sglang

[AMD] Support alt stream for Qwen3.5 on AMD platform

合并时间：2026-06-05 19:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25885

执行摘要

- 一句话：AMD Qwen3.5 alt stream 支持与性能调优
- 推荐动作：值得精读，尤其是如何通过环境变量和 server args 精细控制子模块行为，并在性能与兼容性之间做出权衡。设计思路可推广到其他模型的类似优化。

功能与动机

在 AMD 平台上启用 CUDA alt stream 可以重叠计算以提升推理性能，但直接全量启用会导致 QK norm 和 GDN 模块的性能回归，且与已有的共享专家融合机制产生冲突。因此需要精细控制各子模块的 alt stream 开关，并根据共享专家融合状态自动启用或禁用。

实现拆解

1. 模块级环境变量缓存：在文件 python/sglang/srt/models/qwen3_5.py 模块顶部新增 _hip_use_alt_stream、_gdn_use_alt_stream、_qknorm_use_alt_stream 三个变量，分别对应 SGLANG_ALT_STREAM、SGLANG_GDN_QKVZ_BA_ALT_STREAM、SGLANG_QK_NORM_ALT_STREAM 环境变量与 _is_hip 的与运算结果，避免在热路径中重复读取环境变量。
2. 共享专家融合状态检测：新增函数 _disable_shared_experts_fusion() 延迟从全局 server args 读取 --disable-shared-experts-fusion 标志，用于决定 MoE 层是否启用 alt stream 和共享专家融合。
3. alt stream 创建条件扩展：在 Qwen3_5ForCausalLM.__init__ 中，将 alt stream 创建条件从 _is_cuda 扩展为 _is_cuda or _hip_use_alt_stream，使 AMD 平台也能创建 CUDA stream。
4. GDN 输入投影使用 alt stream 条件：在 Qwen3_5GatedDeltaNet._forward_input_proj 的 alt stream 分支后添加 and _gdn_use_alt_stream，默认不启用 GDN 部分的 alt stream。
5. QK 归一化使用 alt stream 条件：在 _apply_qk_norm 中添加 and _qknorm_use_alt_stream，默认不启用 QK norm 的 alt stream。
6. MoE 层 alt stream 与共享专家融合协调：在 Qwen3_5DecoderLayer.__init__ 中，根据 _disable_shared_experts_fusion() 决定传入 MoE 块的 alt_stream 参数和 support_shared_expert_fusion 参数；当共享专家融合启用时（disable_shared_experts_fusion=False），alt_stream 设为 None 以自动禁用 alt stream。

7. num_fused_shared_experts 清零：在 Qwen3_5ForCausalLM.__init__ 中，当 _use_aiter 且 not _disable_shared_experts_fusion() 时将 num_fused_shared_experts 置为 0，确保共享专家融合与 alt stream 互斥。

关键文件：

- python/sglang/srt/models/qwen3_5.py（模块 模型层；类别 source；类型 core-logic；符号 _disable_shared_experts_fusion, _hip_use_alt_stream, _gdn_use_alt_stream, _qknorm_use_alt_stream）：唯一修改的文件，包含所有 alt stream 控制逻辑：环境变量读取、共享专家融合检测、GDN/QK norm/MoE 层的 alt stream 条件修改。

关键符号：_disable_shared_experts_fusion, Qwen3_5GatedDeltaNet._forward_input_proj, Qwen3_5DecoderLayer.init, Qwen3_5ForCausalLM.init, _apply_qk_norm

关键源码片段

python/sglang/srt/models/qwen3_5.py

唯一修改的文件，包含所有 alt stream 控制逻辑：环境变量读取、共享专家融合检测、GDN/QK norm/MoE 层的 alt stream 条件修改。

```
# python/sglang/srt/models/qwen3_5.py

# === 模块级 alt stream 控制变量 ===
# 读取环境变量并缓存，避免热路径重复调用 get_bool_env_var
_hip_use_alt_stream = get_bool_env_var("SGLANG_ALT_STREAM") and _is_hip
_gdn_use_alt_stream = (
    get_bool_env_var("SGLANG_GDN_QKVZ_BA_ALT_STREAM", "False") and _hip_use_alt_stream
)
_qknorm_use_alt_stream = (
    get_bool_env_var("SGLANG_QK_NORM_ALT_STREAM", "False") and _hip_use_alt_stream
)

# 延迟获取 server args，避免模块导入时访问未初始化的全局状态
def _disable_shared_experts_fusion() -> bool:
    # server args 在 import 时可能还未设置（例如单元测试）
    return get_global_server_args().disable_shared_experts_fusion

# === GDN 输入投影中使用 alt stream ===
# 仅当 alt stream 存在、处于捕获模式、序列长度小于阈值且 _gdn_use_alt_stream 为 True 时启用
if (
    self.alt_stream is not None
    and get_is_capture_mode()
    and seq_len < DUAL_STREAM_TOKEN_THRESHOLD
    and _gdn_use_alt_stream # 新增条件，默认不启用
):
    current_stream = torch.cuda.current_stream()
    self.alt_stream.wait_stream(current_stream)
    projected_states_qkvz, _ = self.in_proj_qkvz(hidden_states)
    with torch.cuda.stream(self.alt_stream):
```

```
            projected_states_ba, _ = self.in_proj_ba(hidden_states)
        current_stream.wait_stream(self.alt_stream)
    else:
        projected_states_qkvz, _ = self.in_proj_qkvz(hidden_states)
        projected_states_ba, _ = self.in_proj_ba(hidden_states)

# === MoE 层初始化：根据共享专家融合状态决定 alt stream ===
# 当共享专家融合启用时（disable_shared_experts_fusion=False），
# alt stream 设为 None 以自动禁用；反之则传入外部的 alt_stream
Qwen2MoeSparseMoeBlock(
    layer_id=layer_id,
    config=config,
    quant_config=quant_config,
    alt_stream=(alt_stream if _disable_shared_experts_fusion() else None),
    prefix=...,
    is_nextn=is_nextn,
    support_shared_expert_fusion=not _disable_shared_experts_fusion(),
)

# === QK 归一化中使用 alt stream 条件 ===
# 同样默认不启用 (_qknorm_use_alt_stream 默认为 False)
if (
    self.alt_stream is not None
    and get_is_capture_mode()
    and _qknorm_use_alt_stream # 新增条件
):
    current_stream = torch.cuda.current_stream()
    self.alt_stream.wait_stream(current_stream)
    q_by_head = q.reshape(-1, self.head_dim)
    with torch.cuda.stream(self.alt_stream):
        # QK norm 在 alt stream 上执行
        q = self.q_norm(q_by_head).reshape(-1, num_kv_groups, self.head_dim)
        k = self.k_norm(k.reshape(-1, self.head_dim)).reshape(-1, num_kv_groups, self.head_dim)
    current_stream.wait_stream(self.alt_stream)
else:
    q = self.q_norm(q.reshape(-1, self.head_dim)).reshape(-1, num_kv_groups, self.head_dim)
    k = self.k_norm(k.reshape(-1, self.head_dim)).reshape(-1, num_kv_groups, self.head_dim)

# === alt stream 创建条件拓展 ===
# 之前仅对 CUDA 创建 stream，现在 AMD HIP 也可创建
alt_stream = torch.cuda.Stream() if (_is_cuda or _hip_use_alt_stream) else None
```

评论区精华

- 环境变量缓存建议：gemini-code-assist 建议将 SGLANG_QK_NORM_ALT_STREAM 和 SGLANG_GDN_QKVZ_BA_ALT_STREAM 缓存到模块级变量，避免热路径重复调用 get_bool_env_var，作者采纳并实现。

- 共享专家融合控制复用：kkHuang-amd 指出应使用已有的 server args --disable-shared-experts-fusion 而非新增环境变量 SGLANG_MOE_SHARED_EXPERT_FUSION，作者随后移除环境变量并改用 server args。
- 性能权衡讨论：hubertlu-tw 与作者深入讨论了小 batch 下 alt stream 的 async EP 开销可能超过收益，作者补充了多并发测试数据，确认 alt stream 在 QKNorm 和 GDN 上存在回归，因此默认关闭。
- CI 修复：初始 CI 失败，作者通过将 _disable_shared_experts_fusion 改为函数延迟调用修复了模块导入时的 server args 未初始化问题。
  - 环境变量缓存建议 (performance): 作者采纳，添加了 _gdn_use_alt_stream 和 _qknorm_use_alt_stream 变量。
  - 共享专家融合控制复用 (design): 作者移除环境变量，改用 server args，新增函数 _disable_shared_experts_fusion()。
  - alt stream 性能权衡讨论 (performance): 决定默认关闭 QKNorm 和 GDN 的 alt stream，仅保留 MoE 部分的 alt stream 控制。
  - CI 错误修复 (correctness): 修复后 CI 通过。

风险与影响

- 风险：
  1. 环境变量依赖风险：如果用户未设置 GPU_MAX_HW_QUEUES（建议至少 5），alt stream 可能无法正常工作或性能下降。
  2. 性能回归风险：尽管默认关闭了 QKNorm 和 GDN 的 alt stream，但用户在启用后可能遇到性能下降，尤其是小 batch 场景。
  3. 配置冲突风险：当 disable_shared_experts_fusion=False 时强制禁用 alt stream，若用户同时启用两者可能无法达到预期效果，但已通过代码逻辑自动规避。
  4. 仅 AMD 平台生效：通过 _is_hip 隔离，不影响其他平台，但引入的代码复杂度增加维护成本。 - 影响：影响范围：仅限 AMD 平台运行 Qwen3.5 模型的场景。影响程度：在禁用共享专家融合时，开启 alt stream 可获得约 1.76% 的 TPOT 提升；若已启用共享专家融合（性能提升约 16.6%），则 alt stream 自动关闭，无额外影响。高并发下趋势一致。对 NVIDIA 和其他平台无影响。测试覆盖：未新增单元测试，但提供了详细的精度和性能 benchmark 数据（GSM8K 精度 0.955，多并发 TPOT 表格）。
- 风险标记：AMD 平台限定，环境变量依赖，性能回归风险，与共享专家融合冲突

关联脉络

- 暂无明显关联 PR