PR #34938 完整报告

sgl-project/sglang

perf: overlap Qwen shared expert with DeepEP routed experts

合并时间：2026-08-22 06:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34938

执行摘要

- 一句话：Qwen 共享专家与 DeepEP 路由专家重叠，吞吐提升约 2%
- 推荐动作：值得精读，特别是对 MoE 性能优化感兴趣的工程师。关键设计决策包括：使用备用流重叠共享专家与路由专家，通过事件实现同步，以及针对可断点 CUDA 图的回退策略。建议关注后续是否补充单元测试。

功能与动机

PR body 指出：对于使用 CUDA 和 DeepEP 的 Qwen MoE 模型，共享专家当前在默认流上于路由专家路径之前运行，导致两个独立计算被串行化，备用流被闲置。优化目标是利用备用流重叠这两个计算，减少解码延迟并提升吞吐。

实现拆解

1. 新增环境变量：在 python/sglang/srt/environ.py 的 Envs 类中添加 SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP，默认值为 True（提交 f55be16795 将默认从 False 改为 True，并在后续重命名为最终名称）。
2. 实现重叠逻辑：在 python/sglang/srt/models/qwen2_moe.py 的 _forward_deepep 方法中增加 enable_cuda_shared_overlap 条件判断，当 CUDA、启用标志、非可断点 CUDA 图、存在 alt_stream 和 shared_expert 且 token 数大于 0 时，将共享专家计算派发到 alt_stream，并记录事件；在主流的 DeepEP 路由专家执行完成后，通过 current_stream.wait_event(shared_event) 等待共享专家完成，再累加 shared_output。为避免可断点 CUDA 图跨段重叠问题，在该场景下保留原有串行路径。
3. 更新文档：在 docs/docs/references/environment_variables.mdx 中新增环境变量说明，描述其功能和默认值。

关键文件：

- python/sglang/srt/models/qwen2_moe.py（模块 模型执行；类别 source；类型 core-logic；符号 _forward_deepep）：核心实现文件，在 _forward_deepep 中新增重叠执行逻辑。
- python/sglang/srt/environ.py（模块 配置中心；类别 source；类型 configuration；符号 SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP）：新增环境变量定义，控制优化开关。

- docs/docs/references/environment_variables.mdx（模块 文档；类别 docs；类型 documentation）：文档更新，说明新环境变量的用途和默认值。

关键符号：_forward_deepep

关键源码片段

python/sglang/srt/models/qwen2_moe.py

核心实现文件，在 _forward_deepep 中新增重叠执行逻辑。

```
# python/sglang/srt/models/qwen2_moe.py 中的关键修改片段
# 在 _forward_deepep 方法内，新增 enable_cuda_shared_overlap 条件，
# 用于判断是否启用 CUDA 备用流重叠优化。
def _forward_deepep(self, hidden_states, forward_batch):
    # 原有 NPU 双流逻辑
    enable_dual_stream = (
        is_npu()
        and envs.SGLANG_NPU_USE_MULTI_STREAM.get()
        and forward_batch.forward_mode.is_cuda_graph()
    )

    # 新增 CUDA 重叠条件：仅在 CUDA、启用环境变量、非可断点 CUDA 图、
    # 存在 alt_stream 和 shared_expert 且 token 数大于 0 时生效。
    enable_cuda_shared_overlap = (
        _is_cuda
        and envs.SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP.get()
        # 可断点 CUDA 图会在 eager DeepEP break 前合并侧流，
        # 因此该路径无法实现两个专家计算的真正重叠。
        and not is_in_breakable_cuda_graph()
        and self.alt_stream is not None
        and self.shared_expert is not None
        and hidden_states.shape[0] > 0
    )

    shared_output = None
    if hidden_states.shape[0] > 0:
        # 路由 logits 计算
        router_logits, _ = self.gate(hidden_states)
        if enable_dual_stream:
            # NPU 双流路径（原有）
            shared_output = shared_expert_on_independent_stream(
                hidden_states.clone(), self._forward_shared_experts
            )
        elif enable_cuda_shared_overlap:
            # CUDA 备用流重叠路径：将共享专家派发到 alt_stream，
            # 主流继续执行 DeepEP 路由专家，最后通过事件同步。
            current_stream = torch.cuda.current_stream()
            self.alt_stream.wait_stream(current_stream)
            with torch.cuda.stream(self.alt_stream):
```

```
                shared_output = self._forward_shared_experts(hidden_states)
                shared_output.record_stream(self.alt_stream)
                shared_event = self.alt_stream.record_event()
        else:
            # 原有串行路径
            shared_output = self._forward_shared_experts(hidden_states)

        # 路由专家处理（DeepEP）
        topk_output = self.topk(...)

    final_hidden_states = self.experts(hidden_states, topk_output=topk_output)

    # 等待备用流完成，再累加共享专家输出
    if enable_dual_stream:
        wait_share_stream()
    elif enable_cuda_shared_overlap:
        torch.cuda.current_stream().wait_event(shared_event)

    if shared_output is not None:
        final_hidden_states.add_(shared_output)

    return final_hidden_states
```

python/sglang/srt/environ.py

新增环境变量定义，控制优化开关。

```
# python/sglang/srt/environ.py 中的 Envs 类片段
# 在 MoE 执行相关配置区新增环境变量，默认启用重叠优化。
class Envs:
    # ...
    SGLANG_BLACKWELL_OVERLAP_SHARED_EXPERTS_OUTSIDE_SBO = EnvBool(False)
    # 新增：启用 Qwen DeepEP 共享专家重叠优化，默认开启。
    SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP = EnvBool(True)
    # ...
```

评论区精华

1. nvpohanh 提问：若总是有益，是否应默认启用？作者回应已在 f55be16795 中默认启用。
2. ch-wan 建议：将变量名重命名为 SGLANG_OPT_QWEN_DEEPEP_SHARED_OVERLAP 或 SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP，并同步更新文档和调用点。作者已按建议重命名为 SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP 并更新所有引用。

- 是否默认启用 (design): 作者确认已在 f55be16795 中默认启用。
- 环境变量命名 (style): 作者已重命名为 SGLANG_ENABLE_QWEN_DEEPEP_SHARED_OVERLAP 并更新所有引用。

风险与影响

- 风险：
  1. 正确性风险：重叠执行涉及多流同步，若事件等待缺失或流顺序错误，可能产生数据竞争或错误结果。虽然作者提供了精度测试（GSM8K 满分 200/200），但未覆盖所有模型和负载。
  2. 性能回退风险：默认启用后，若特定硬件或配置下重叠收益不显著，可能引入额外流切换开销，反而降低性能。
  3. 可断点 CUDA 图兼容性：已排除该优化，但仍需关注其他 CUDA 图模式下的行为。
  4. 测试覆盖不足：PR 未添加单元测试，依赖端到端验证，长期维护风险较高。 - 影响：影响范围限于 CUDA 上使用 DeepEP 的 Qwen MoE 模型（如 Qwen3.5-397B-A17B）。对用户而言，默认启用后可能获得约 2% 的吞吐提升和约 3% 的延迟降低，无需额外配置。对系统而言，新增一个环境变量和一条重叠路径，对非 Qwen 或非 DeepEP 模型无影响。对团队而言，该优化可作为其他 MoE 模型重叠优化的参考。 - 风险标记：核心路径变更，缺少测试覆盖，与 CUDA 图交互

关联脉络

- PR #36219 [Performance] Tune FlashInfer EXTEND for DP prefill: 同为性能优化，涉及调度器流调优，与本 PR 的流重叠思路相关。
- PR #36237 [MegaMoE] Respect padded MXFP8 scale row strides in pre-dispatch: 均为 MoE 专家调度 / 执行优化，涉及 DeepEP 相关路径。