PR #23979 完整报告

sgl-project/sglang

Enable DeepGEMM PDL on by default

合并时间：2026-06-05 05:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23979

执行摘要

- 一句话：默认启用 DeepGEMM PDL
- 推荐动作：此 PR 是低风险、有明确性能收益的微小优化，适合合并。建议相关工程师了解 PDL 的基本原理，以及通过环境变量控制该特性的方式。

功能与动机

DeepGEMM 的 Programmatic Dependent Launch (PDL) 功能可以优化内核启动流水线，但此前是 opt-in 状态，需要用户手动调用 deep_gemm.set_pdl(True)。为了在所有场景下默认获得性能收益（GQPA 评测显示约 0.8% 的延迟改善），需要将 PDL 默认启用。PR 中作者通过 GLM-5 FP8 的 GPQA 评测展示了性能提升，并指出 deep_gemm.set_pdl 的检查是安全的，因为即使 deep_gemm 模块没有 set_pdl 导出也不会报错。

实现拆解

1. 环境变量配置 (environ.py)：在 Envs 类中新增 SGLANG_DEEPGEMM_PDL = EnvBool(True)，默认值为 True，放在 DeepGemm 环境变量区块中。
2. 入口点调用 (entrypoint.py)：在 if ENABLE_JIT_DEEPGEMM: 代码块内部，即导入 deep_gemm 之后，添加条件判断：如果 envs.SGLANG_DEEPGEMM_PDL.get() 为真且 deep_gemm 模块包含 set_pdl 属性，则调用 deep_gemm.set_pdl(True)。由于该调用发生在导入后的模块作用域中，所有后续 DeepGEMM 内核调用都将自动使用 PDL 模式。
3. 向后兼容：通过 hasattr(deep_gemm, "set_pdl") 确保即使 deep_gemm 版本较老、不包含 set_pdl 函数，代码也不会崩溃。

关键文件：

- python/sglang/srt/layers/deep_gemm_wrapper/entrypoint.py（模块 内核调度；类别 source；类型 core-logic）：DeepGEMM 内核的入口点，新增了 PDL 全局启用的调用逻辑。
- python/sglang/srt/environ.py（模块 配置层；类别 source；类型 core-logic）：定义了新的环境变量 SGLANG_DEEPGEMM_PDL，默认值为 True。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/deep_gemm_wrapper/entrypoint.py

DeepGEMM 内核的入口点，新增了 PDL 全局启用的调用逻辑。

```
# python/sglang/srt/layers/deep_gemm_wrapper/entrypoint.py
# 在 deep_gemm 导入后立即启用 PDL 特性
if ENABLE_JIT_DEEPGEMM:
    import deep_gemm
    from deep_gemm.utils.layout import get_mn_major_tma_aligned_tensor # noqa: F401

    # 如果环境变量 SGLANG_DEEPGEMM_PDL 为 True，
    # 并且当前 deep_gemm 版本支持 set_pdl 函数，
    # 则全局启用 Programmatic Dependent Launch (PDL) 模式
    if envs.SGLANG_DEEPGEMM_PDL.get() and hasattr(deep_gemm, "set_pdl"):
        deep_gemm.set_pdl(True)
```

python/sglang/srt/environ.py

定义了新的环境变量 SGLANG_DEEPGEMM_PDL，默认值为 True。

```
# python/sglang/srt/environ.py
# DeepGemm 配置区块中新增 PDL 开关，默认启用
class Envs:
    # ...
    # DeepGemm
    SGLANG_ENABLE_JIT_DEEPGEMM = EnvBool(True)
    SGLANG_JIT_DEEPGEMM_PRECOMPILE = EnvBool(True)
    SGLANG_JIT_DEEPGEMM_FAST_WARMUP = EnvBool(False)
    SGLANG_JIT_DEEPGEMM_COMPILE_WORKERS = EnvInt(4)
    SGLANG_IN_DEEPGEMM_PRECOMPILE_STAGE = EnvBool(False)
    SGLANG_DG_CACHE_DIR = EnvStr(os.path.expanduser("~/.cache/deep_gemm"))
    SGLANG_DG_USE_NVRTC = EnvBool(False)
    SGLANG_USE_DEEPGEMM_BMM = EnvBool(False)
    SGLANG_DEEPGEMM_SANITY_CHECK = EnvBool(False)
    SGLANG_DEEPGEMM_PDL = EnvBool(True) # 新增：默认启用 PDL
    SGLANG_PP_PARALLEL_DEEPGEMM_WARMUP = EnvBool(False)
    # ...
```

评论区精华

无实质性 review 讨论。只有一个审核者 Fridge003 批准了 PR，无评论。作者在 PR 正文中说明了性能提升（约 0.8%）以及安全性（set_pdl 检查不会导致错误）。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险：低。即使 deep_gemm 版本不支持 PDL，hasattr 检查会跳过调用，行为不变。PR 默认开启 PDL，但用户可通过设置环境变量 SGLANG_DEEPGEMM_PDL=0 关闭，回退容易。

    - 性能风险：根据作者提供的 GPQA 评测（重复 8 次，mean 0.850，平均延迟 12683s），PDL 带来了约 0.8% 的性能提升，无明显负面性能影响。
    - 兼容性风险：低。仅依赖 deep_gemm.set_pdl 的存在性检查，不影响其他模块。
    - 安全风险：无。
- 影响：
    - 用户影响：所有使用 DeepGEMM 内核（GEMM，MoE 等）的用户将自动获得约 0.8% 的延迟改善，无需任何配置更改。
    - 系统影响：PDL 特性影响所有基于 SM90 和 SM100 的 DeepGEMM 调用，可能对 MoE 调度和整体吞吐量产生正向影响。
    - 团队影响：无。
    - 风险标记：配置默认值变更

关联脉络

- PR #25239 [FlashInfer v0.6.12] Support FlashInfer 4over6 NVFP4: 同属低精度计算性能优化方向，涉及 DeepGEMM 相关内核和性能提升。
- PR #27111 [AMD] Minimax M25：FP8 block-scale GEMM dispatch for ROCm 7.0 on gfx950: 同为 FP8 GEMM 性能优化，虽面向 AMD 平台，但属于相同的 DeepGEMM 技术栈。