PR #24815 完整报告

sgl-project/sglang

Revert "[NPU] fix profiler on npu"

合并时间：2026-05-09 17:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24815

执行摘要

- 一句话：回滚 NPU 分析器修复 PR#24685
- 推荐动作：该 PR 为简单回滚，无需深入审查。建议关注后续是否有重新修复 NPU profiler 的 PR，以更 robust 的方式传入 experimental_config。

功能与动机

这是一次自动回滚操作，引用 PR#24685 的具体动机：修复 sglang.bench_serving --profile 在 NPU 上无法收集算子形状信息的问题，原 PR 增加了 experimental_config 参数。回滚原因未在 body 中说明，可能由于该参数导致未知问题。

实现拆解

1. 定位文件 python/sglang/srt/managers/scheduler_profiler_mixin.py 中 start_profile() 方法（约第 192-204 行）。
2. 删除 torch.profiler.profile() 调用中的 experimental_config 关键字参数及其条件表达式 None if not _is_npu else torch_npu.profiler._ExperimentalConfig(...)，共 15 行。
3. 调用 self.torch_profiler.start() 保持不变，后续 profiling 流程未受影响。

关键文件：

- python/sglang/srt/managers/scheduler_profiler_mixin.py（模块 分析器；类别 source；类型 core-logic）：唯一变更文件，删除了 NPU 特定的 experimental_config 参数，直接影响 NPU profiling 行为。

关键符号：start_profile

关键源码片段

python/sglang/srt/managers/scheduler_profiler_mixin.py

唯一变更文件，删除了 NPU 特定的 experimental_config 参数，直接影响 NPU profiling 行为。

```
def start_profile(self: Scheduler, req: ProfileReq) -> ProfileReqOutput:
    # ... 前面的逻辑 ...
    elif torchprof_activities:
        self.torch_profiler = torch.profiler.profile(
```

```
            activities=torchprof_activities,
            with_stack=with_stack if with_stack is not None else True,
            record_shapes=record_shapes if record_shapes is not None else False,
            on_trace_ready=(
                None
                if not _is_npu
                else torch_npu.profiler.tensorboard_trace_handler(
                    str(self.torch_profiler_output_dir)
                )
            ),
            # 删除了 experimental_config 参数，详见 PR#24685 回滚
        )
        self.torch_profiler.start()
        self.profile_in_progress = True
    # ... 后续逻辑 ...
```

评论区精华

无实质性讨论。只有 gemini-code-assist[bot] 的自动评论，未产生 human review comments。

- 暂无高价值评论线程

风险与影响

- 风险：回滚后，NPU 平台将失去 experimental_config 中的定制设置（如 profiler_level=Level1、aic_metrics=AiCoreNone 等），可能导致 NPU 上 profiling 信息不完整，无法收集算子形状，恢复修复前的状态。不过由于该 PR 本身是对修复的回滚，如原修复未彻底问题则回滚反而稳定。
- 影响：仅影响 NPU 平台上的 profiling 功能。用户在使用 sglang.bench_serving --profile 时，若在 NPU 环境，无法获得算子形状等详细信息；其他平台（CUDA、ROCM）无影响。
- 风险标记：NPU 平台 profiling 退化，无 review 讨论

关联脉络

- PR #24685 [NPU] fix profiler on npu: 被回滚的原始 PR，其内容完全被撤销。