

PR #28165 完整报告

sgl-project/sglang

Unify NVTX annotation helpers and split the enable gate per subsystem

合并时间: 2026-06-14 15:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28165>

执行摘要

- 一句话: 统一 profiler 注释, 按子系统拆分 NVTX 门控
- 推荐动作: 值得所有关注 SGLang 性能分析能力的开发者精读。本 PR 展示了一个优秀的重构案例: 如何通过定义统一原语 (profile_range/profile_method) 并利用偏函数预绑定子系统门控, 将分散的 profiling 逻辑集中化。对于希望扩展 SGLang profiling 的开发者, 这是必读的基础设施变更。

功能与动机

SGLang 中存在多个分散的 profiler span 实现: nvtx_utils 仅用于调度器且要求 nvtx 包; batch_overlap/operations.py 有独立的 _annotate_region 并从 os.environ 直接读取 SGLANG_OPERATIONS_ENABLE_PROFILE; spec_stage_span 和 model_forward step span 使用第三种方式仅 emit record_function 且门控依赖 torch profiler 是否激活。这导致两个问题:

1. record_function 与 NVTX 耦合, 使得在未安装 nvtx 包时调度器和 batch-overlap 的 span 在纯 torch.profiler 捕获中不可见。
2. 门控与环境变量命名不一致, 且 SGLANG_OPERATIONS_ENABLE_PROFILE 未在 environ.py 中注册。本 PR 致力于统一这些实现, 解耦发射器, 并标准化门控命名。

实现拆解

该 PR 按以下步骤完成统一:

1. 重写 nvtx_utils.py: 定义共享原语 profile_range (上下文管理器) 和 profile_method (装饰器)。内部使用 _profile_range_impl 根据条件选择性地发射 record_function (当 torch profiler 活动时) 和 NVTX range (当 per-subsystem nvtx_enabled 为 True 时)。同时提供预绑定的辅助函数 scheduler_nvtx_method 和 operations_nvtx_range。
2. 注册新环境变量: 在 environ.py 中添加 SGLANG_ENABLE_NVTX_SCHEDULER 和 SGLANG_ENABLE_NVTX_OPERATIONS, 并保留旧名称 SGLANG_ENABLE_NVTX 和 SGLANG_OPERATIONS_ENABLE_PROFILE 作为弃用别名。
3. 统一消费者:
 - scheduler.py 和 request_receiver.py: 将 @nvtx_annotated_method 替换为 @scheduler_nvtx_method。

- `batch_overlap/operations.py`: 删除独立的 `_annotate_region` 上下文管理器及其引入的 `os.environ` 读取、`nvtx` 导入, 改用 `operations_nvtx_range`。
- `spec_utils.py`: `spec_stage_span` 函数重写为直接调用 `profile_range`。
- `model_runner.py`: `forward` 方法中的 `step span` 从条件式 `torch.profiler.record_function` 替换为 `profile_range`。
- `disaggregation/decode.py` 和 `disaggregation/prefill.py`: 更新导入使其使用新的 `profile_range` 符号。

4. 测试与配置: 本次改动主要涉及源码重构, 未添加新测试, 现有 CI 应覆盖功能。

关键文件:

- `python/sglang/srt/utils/nvtx_utils.py` (模块 NVTX 工具; 类别 source; 类型 core-logic; 符号 `_profile_range_impl`, `profile_range`, `profile_method`, `scheduler_nvtx_method`): 重构核心文件, 定义了统一的 profiler span 原语 `profile_range/profile_method`, 解耦了 `record_function` 和 NVTX range 发射器, 并实现了 per-subsystem 门控。所有后续消费者依赖于此。
- `python/sglang/srt/batch_overlap/operations.py` (模块 批处理重叠; 类别 source; 类型 core-logic; 符号 `_annotate_region`): 移除了独立的 `_annotate_region` 实现及其直接 `os.environ` 读取, 改用统一的 `operations_nvtx_range`, 展示了新原语在 batch-overlap 子系统中的应用。
- `python/sglang/srt/model_executor/model_runner.py` (模块 模型运行器; 类别 source; 类型 data-contract): `forward` 方法中的 `step span` 从条件式 `torch.profiler.record_function` 替换为统一的 `profile_range`, 展示了该原语在核心 `forward` 路径上的应用。
- `python/sglang/srt/managers/scheduler.py` (模块 调度器; 类别 source; 类型 dependency-wiring): 所有调度器主要方法 (`process_input_requests`, `get_next_batch_to_run`, `run_batch`, `process_batch_result`) 的装饰器从 `nvtx_annotated_method` 切换为 `scheduler_nvtx_method`, 体现了调度子系统门控的应用。
- `python/sglang/srt/speculative/spec_utils.py` (模块 推测解码; 类别 source; 类型 dependency-wiring): `spec_stage_span` 函数从条件返回 `record_function/nullcontext` 改为直接调用 `profile_range`, 统一了推测解码阶段的 span。
- `python/sglang/srt/envron.py` (模块 环境变量; 类别 source; 类型 core-logic): 注册了新的环境变量 `SGLANG_ENABLE_NVTX_SCHEDULER` 和 `SGLANG_ENABLE_NVTX_OPERATIONS`, 并保留了旧名称作为弃用别名。这是门控配置的中央注册点。
- `python/sglang/srt/managers/scheduler_components/request_receiver.py` (模块 请求接收器; 类别 source; 类型 dependency-wiring): `recv_requests` 方法的装饰器从 `nvtx_annotated_method` 切换为 `scheduler_nvtx_method`。
- `python/sglang/srt/disaggregation/decode.py` (模块 解聚; 类别 source; 类型 dependency-wiring): 更新导入以使用新的 `profile_range` 符号, 保持解聚解码阶段的注释一致性。

- python/sglang/srt/disaggregation/prefill.py (模块 解聚; 类别 source; 类型 dependency-wiring) : 更新导入以使用新的 profile_range 符号, 保持解聚预填充阶段的注释一致性。

关键符号: profile_range, profile_method, _profile_range_impl, scheduler_nvtx_method, operations_nvtx_range

关键源码片段

python/sglang/srt/batch_overlap/operations.py

移除了独立的 _annotate_region 实现及其直接 os.environ 读取, 改用统一的 operations_nvtx_range, 展示了新原语在 batch-overlap 子系统中的应用。

```
# 在 _StageExecutor.next 方法中, 原使用 _annotate_region 自定义上下文管理器
# 现在直接使用从 nvtx_utils 导入的 operations_nvtx_range (预绑定了 NVTX_OPERATIONS_ENABLED 门控)
from sglang.srt.utils.nvtx_utils import operations_nvtx_range
```

```
def next(self):
    assert not self.done
    stage = self._stages[self._index]
    # ... set_dp_buffer_len ...
    ctx_mgr = (
        forward_context(self._child_ctx)
        if self._child_ctx is not None
        else nullcontext()
    )
    # 用预绑定的 operations_nvtx_range 替换独立 _annotate_region
    stage_range = operations_nvtx_range(
        debug_name=f'{self._debug_name}{self._index}',
        color='orange',
    )
    with ctx_mgr, stage_range:
        for op in stage:
            # 每个操作也使用 operations_nvtx_range, 并指定颜色为黄色
            with operations_nvtx_range(
                debug_name=op.debug_name,
                color='yellow',
            ):
                self._stage_output = op.fn(
                    state=self._stage_state,
                    **(
                        self._stage_output if self._stage_output is not None else {}
                    ),
                )
    self._index += 1
```

评论区精华

Review 提出了两点中等优先级的改进建议：

- 在 `nvtx_utils.py` 中，当 `nvtx` 包缺失时打印警告应仅在 `rank 0` 进程进行，以避免多 GPU 环境下日志重复。评论建议添加 `os.getenv("RANK", "0") == "0"` 的检查。
- 在 `nvtx` 颜色参数处理中，应使用 `if color:` 而非 `color is not None`，以安全处理空字符串，避免将空字符串传递给 `nvtx` 注释。两点建议均未被合并前的版本采纳（PR 已合并但未应用修改）。
- 多 GPU 环境下日志去重 (other): 建议未在合并版本中采纳（PR 已合并但无 `rank` 过滤）。
- `color` 参数的空字符串安全性 (design): 建议未在合并版本中采纳（仍使用 `color is None` 的检查）。

风险与影响

- 风险：总体风险较低。主要考量：
 - 向后兼容：旧环境变量 `SGLANG_ENABLE_NVTX` 和 `SGLANG_OPERATIONS_ENABLE_PROFILE` 仍有效（但触发 `DeprecationWarning`），因此现有启动脚本不受影响。
 - 缺失消费者：如有未覆盖的注释站点仍使用旧的独立实现（如 `nvtx_pytorch_hooks.py` 未改动），但该模块并非本次统一目标，其行为保持不变。
 - 默认行为变更：之前部分 `span`（如 `scheduler` 的 `NVTX range`）需要显式设置 `SGLANG_ENABLE_NVTX` 才记录；现在 `span` 默认通过 `record_function` 在 `torch profiler` 活跃时可见。这可能导致在未见过的 `profiler` 输出中看到新 `span`，但这是期望的改进。
 - 多 GPU 日志：未采纳 `rank 0` 日志过滤建议，因此多进程场景下仍会有重复缺失警告，但不影响功能。
 - 影响：用户：无需更改配置。原 `SGLANG_ENABLE_NVTX` 用户仍可继续使用，但建议迁移到新名称。`torch.profiler` 用户会发现更多 `span`（调度器、推测解码、前向）自动出现，无需设置环境变量。系统：统一了 `profiler` 基础设施，消灭了代码重复，降低了未来维护成本。每个子系统的 `NVTX` 门控独立，可以单独开启调度器或 `batch-overlap` 的 `Nsight` 标记。团队：核心贡献者 `hnyls2002` 此前在 PR #27901 中为调度器添加了 `NVTX` 标记，本次在此基础上推广到所有子系统并解耦发射器。
- 风险标记：统一重构可能导致遗漏消费者，环境变量重命名需向后兼容，默认行为变更：`span` 现在默认通过 `record_function` 可见，未采纳 `rank 0` 日志过滤

关联脉络

- PR #27901 feat: add NVTX markers for the scheduler main loop: 本 PR 的作者 `hnyls2002` 在 PR #27901 中为调度器主循环首次添加了 `NVTX markers`，本 PR 在此基础上将 `NVTX markers` 推广至所有子系统并解耦 `record_function` 和 `NVTX` 发射器。