

PR #36149 完整报告

sgl-project/sglang

fix(xpu): read enable_deterministic_inference from the config bag

合并时间: 2026-08-24 13:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36149>

执行摘要

- 一句话: 修复 XPU 后端从配置袋读取确定性推理开关
- 推荐动作: 值得精读, 因为这是一个典型的配置袋迁移遗留问题的修复, 展示了如何在 SGLang 中正确读取解析后的配置值。对于维护注意力后端或涉及确定性推理的工程师很有参考价值。

功能与动机

test/registered/unit/test_chain_read_ratchet.py 在 main 分支上失败, 报错显示 XPUAttentionBackend.__init__ 通过 model_runner.server_args.enable_deterministic_inference 读取字段, 但该路径返回的是原始 CLI 默认值, 而非解析后的值。这源于 #29143 使用的旧拼写与配置袋迁移 (#35907/#35910) 之间的合并竞争。PR 指出其他所有注意力后端都已通过 get_exec().deterministic 读取该字段, XPU 是最后一个遗留点, 也是 srt/ 中唯一对解析字段的借用记录链读取。

实现拆解

1. 在 python/sglang/srt/layers/attention/xpu_backend.py 中, 将 num_splits 的赋值从 model_runner.server_args.enable_deterministic_inference 改为 get_exec().deterministic.enable_deterministic_inference。
2. 同时从 sglang.srt.runtime_context 导入 get_exec, 以支持新的读取方式。
3. 变更仅此一处, 没有新增测试, 因为现有 ratchet 测试即覆盖该行为, 变更后测试转绿。

关键文件:

- python/sglang/srt/layers/attention/xpu_backend.py (模块 注意力后端; 类别 source; 类型 core-logic) : 这是唯一的变更文件, 修复了 num_splits 的配置读取逻辑, 从 model_runner.server_args 改为 get_exec().deterministic, 解决了 ratchet 测试失败并保持与其他后端一致。

关键符号: XPUAttentionBackend.init

关键源码片段

[python/sglang/srt/layers/attention/xpu_backend.py](#)

这是唯一的变更文件，修复了 `num_splits` 的配置读取逻辑，从 `model_runner.server_args` 改为 `get_exec().deterministic`，解决了 `ratchet` 测试失败并保持与其他后端一致。

```
# python/sglang/srt/layers/attention/xpu_backend.py
# 该片段展示了 num_splits 的配置读取逻辑，从旧字段改为配置袋读取。

# 从 runtime_context 导入 get_exec
from sglang.srt.runtime_context import (
    get_exec, # 新增导入，用于访问配置袋
    get_schedule,
    get_spec,
)

# 在 __init__ 中设置 num_splits
self.num_splits = (
    1 if get_exec().deterministic.enable_deterministic_inference else 0
)

# 注释: num_splits == 0 时，内核使用启发式自动确定 split 数量。
# Split-KV 在非确定性分区上执行归约，因此启用确定性推理时固定 num_splits = 1，
# 以保持注意力的归约顺序固定，这与 fa3 后端行为一致。
```

评论区精华

该 PR 无 review 评论。PR 描述中说明了变更的合理性：其他注意力后端（如 `dsa_backend.py:316`、`flashattention_backend.py:320` 等）均已使用 `get_exec().deterministic`，XPU 是最后一个遗留点。CI 中一次 rerun 测试 `test_chain_read_ratchet.py` 成功通过。

- 暂无高价值评论线程

风险与影响

- 风险：本变更仅修改启动时的配置读取路径，不涉及热路径，且默认行为不变（默认关闭时两个拼写结果相同）。但当启用确定性推理时，XPU 行为会从以前可能未固定 `num_splits` 改变为固定为 1，可能带来性能影响（Split-KV 被禁用）。风险较低，但需要确认启用该标志时性能是否符合预期。
- 影响：影响范围限于 XPU 注意力后端，主要影响使用确定性推理的用户。变更确保了确定性推理语义的正确性，修复了 `ratchet` 测试失败。对默认路径无影响。
- 风险标记：配置读取路径变更，可能影响确定性推理性能

关联脉络

- PR #29143 Add intel_xpu to DETERMINISTIC_ATTENTION_BACKEND_CHOICES: 该 PR 添加了 `num_splits` 逻辑，但使用了旧拼写，导致本次修复；本次变更使其从配置袋读取，从而完成该功能的正确实现。
- PR #35907 config bag migration: PR body 提到配置袋迁移（#35907/#35910），本次变更正是为了适配该迁移，使 XPU 后端与其他后端保持一致。