

PR #36222 完整报告

sgl-project/sglang

[CP V1 Deprecation 1/5] Migrate tests to strategy-based prefill CP

合并时间: 2026-08-25 11:15

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36222>

执行摘要

- 一句话: 迁移 prefill CP 测试到新策略参数, 删除 v1 专用测试
- 推荐动作: 值得精读, 作为大型弃用流程中测试先行迁移的参考范式。关注点: 1) 如何在删除旧测试的同时不失去回归保护 (本 PR 依赖后续堆栈及时删除 runtime); 2) 手动测试参数替换的等价性判断; 3) 跨硬件 runner 的验证策略。对 CP 或 DeepSeek 相关测试维护者, 这是一个理解新策略路径入口的窗口。

功能与动机

PR body 明确说明这是 '#35426 替代堆栈的第 1/5 部分', 目标包括 'migrate generic prefill CP tests and manual launch configurations to --enable-prefill-cp --cp-strategy'、'remove explicit SGLANG_ENABLE_CP_V2=1 from allowlisted strategy-path models' 以及 'delete generic v1-only DSA and kernel parity tests'。核心动机是系统化弃用旧的 CP V1 路径, 先让测试基础设施切换到新策略 API, 避免后续删除 runtime 时测试仍然引用旧参数。

实现拆解

1. 手动启动配置迁移: 将 test/manual/dsv4/ 下 7 个测试文件 (B200/B300/GB300/H200 的 flash/pro 变体) 中的 --enable-dsa-prefill-context-parallel + --dsa-prefill-cp-mode round-robin-split 替换为 --enable-prefill-cp + --cp-strategy interleave。
2. 删除 v1-only 测试: 移除 test_mla_cp_fa3_parity.py、test_dsa_prefill_cp_legacy.py、test_cp_prefix_len_fa3_parity.py 三个文件 (合计约 510 行), 这些测试专门钉死旧 SGLANG_ENABLE_CP_V2=0 / DSA 专属路径。
3. 移除显式 env: 在 test_deepseek_v4_flash_fp4_b200_cp.py 中删掉 SGLANG_ENABLE_CP_V2=1 环境变量, 因为新策略路径默认启用 v2。
4. 单元测试收敛: test_cp_strategy_unit.py 中删除 CP v1/v2 双分支验证, 改为只验证 CP v2 分支, 反映新默认行为。
5. 验证配套: PR 自述 test_cp_strategy_unit.py 23 passed、18 subtests; 通过 compileall、pre-commit 注册表校验; AMD/ROCm 与 Ascend NPU 测试树逐字节保持不变; 后续 CI 重跑相关硬件测试全部通过。

关键文件:

- test/registered/kernels/ops/attention/test_mla_cp_fa3_parity.py (模块 注意力测试; 类别 test; 类型 deletion; 符号 _build_cache_and_q, _full_seq_attn, _cp_attn_for_rank,

_mla_cp_attn)：删除的 FA3 MLA prefill CP 数值一致性测试，是 v1 zigzag 路径的 kernel 级回归守护，删除代表旧路径测试覆盖正式移除。

- test/registered/cp/test_dsa_prefill_cp_legacy.py (模块 预填充测试；类别 test；类型 deletion；符号 TestDeepseekV32CPIInSeqSplit, setUpClass, tearDownClass, test_a_gsm8k)：删除的 DeepSeek V3.2 DSA prefill CP 端到端测试，覆盖 in-seq-split 与 round-robin-split 两种旧模式，是 v1 行为的主要回归守护。
- test/registered/kernels/ops/attention/test_cp_prefix_len_fa3_parity.py (模块 前缀并行；类别 test；类型 deletion；符号 TestCPPrefixLenFA3Parity, _run_parity, _call_meta, test_cp2_prefix1_extend3)：删除的 prefix len FA3 parity 测试，专门验证 prepare_context_parallel_metadata 在前缀 + 扩展场景下的 cache_seqlens 正确性，是 v1 元数据路径的关键校验。
- test/registered/cp/test_cp_strategy_unit.py (模块 策略测试；类别 test；类型 test-coverage；符号 test_get_cp_strategy_is_initialized_under_cp_v1_and_cp_v2, test_get_cp_strategy_is_initialized_under_cp_v2)：单元测试从 CP v1/v2 双分支收敛为只验证 v2，反映新默认行为，是本 PR 测试迁移的核心样例。
- test/manual/dsv4/test_b200_flash.py (模块 硬件测试；类别 test；类型 test-coverage)：手动启动配置迁移的代表样例，将 DSA 专属参数替换为通用策略参数，展示新路径的入口调用方式。

关键符号：test_get_cp_strategy_is_initialized_under_cp_v2,
test_enable_cp_v2_and_is_cp_v2_active

关键源码片段

test/registered/cp/test_cp_strategy_unit.py

单元测试从 CP v1/v2 双分支收敛为只验证 v2，反映新默认行为，是本 PR 测试迁移的核心样例。

```
# test/registered/cp/test_cp_strategy_unit.py (修改后关键片段)
# 原有测试同时验证 CP v1 与 CP v2 两个分支；本 PR 起 CP v1 进入弃用流程，
# 因此删除 v1 分支，只保留 v2 作为默认且唯一的验证路径。
def test_get_cp_strategy_is_initialized_under_cp_v2(self):
    # 使用新策略参数初始化：`enable_prefill_cp` 配合 `cp_strategy` 选择 zigzag/interleave
    init_cp_strategy(
        SimpleNamespace(
            enable_prefill_cp=True,
            cp_strategy="interleave",
            attn_cp_size=4,
        )
    )

    # 只保留 `SGLANG_ENABLE_CP_V2=1` 分支，验证新路径下策略对象能正确初始化
    with patch(
        "sglang.srt.environ.envs.SGLANG_ENABLE_CP_V2.get", return_value=True
    ):
        self.assertIsNotNone(get_cp_strategy())
```

```
self.assertTrue(is_cp_enabled())
self.assertTrue(is_interleave())
```

评论区精华

本 PR 没有任何 review 评论或设计讨论。作者 Fridge003 仅在 CI 上使用 `/rerun-group cp` 与 `/rerun-test` 重跑关键硬件测试，github-actions[bot] 返回全部通过。没有公开记录的技术争议或未解决问题。

- CI 重跑验证 (other): github-actions[bot] 反馈所有重跑测试均通过，覆盖 2-gpu-h100、4-gpu-h100、8-gpu-h200、4-gpu-b200 等 runner。

风险与影响

- 风险：
 1. 删除旧路径测试守护：test_dsa_prefill_cp_legacy.py 原本覆盖 DeepSeek V3.2 的 DSA in-seq-split 与 round-robin-split 两种模式，是旧路径的唯一回归守护；删除后若 v1 runtime 尚未移除（堆栈 3/5 才删除），这些路径会失去 CI 覆盖。
 2. 策略语义等价性：迁移后的手动测试改用 `--cp-strategy interleave`，需确认 interleave 语义与原 round-robin-split 等价，否则可能掩盖行为差异；H200/B200 CI 通过降低但不消除该风险。
 3. 环境变量依赖：SGLANG_ENABLE_CP_V2=1 移除后测试默认走 v2，若新策略路径仍依赖旧 env（后续 server args 步骤未完成），存在启动失败风险。
 4. 硬件验证缺口：PR body 注明 GPU 功能测试依赖 H100/H200/B200 匹配 runner，当前 CI 已验证 H200/B200，AMD/NPU 保持不变但未在本 PR 跑新策略。- 影响：对用户和运行时无直接影响，变更全部位于测试目录。对团队的影响是：CI 矩阵减少约 510 行旧路径测试，prefill CP 测试统一收敛到新策略 API，为后续 4 个堆栈步骤（server args、runtime 删除、API 重命名、文档）扫清测试侧障碍。AMD/ROCm 与 Ascend NPU 注册测试保持不变，说明这些平台暂时仍走旧路径，后续步骤需并行关注。- 风险标记：删除旧路径测试守护，依赖后续堆栈步骤，需要硬件 CI 验证

关联脉络

- PR #35426 CP V1 deprecation replacement stack (referenced in PR body): 本 PR 是 #35426 替代堆栈的第 1/5 部分，PR body 明确引用，后续 4 个部分将在此基础上删除 server args 与 runtime。