

PR #37760 完整报告

sgl-project/sglang

[CI][NPU] Fix kimi_k2_6 16p in64k perf test and dsv4-flash testcases

合并时间: 2026-09-03 21:57

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37760>

执行摘要

- 一句话: 修复 NPU 平台上 Kimi K2.6 和 DeepSeek V4-Flash 的测试用例, 优化预填充稳定性并同步至最新硬件配置。
- 推荐动作: 本 PR 是一个重要的测试基础设施维护和同步工作, 值得 NPU 平台开发者和 CI 维护者精读。它揭示了 SGLang 在 NPU 硬件上针对特定模型 (DSV4-Flash) 的性能调优已进入更精细的阶段 (引入 war-barrier、调整内部分块策略)。变更本身风险较低, 但通过阅读可以了解到:
 1. Kimi K2.6 大上下文测试中的一个关键内存优化技巧 (调整 chunked-prefill-size)。
 2. DeepSeek V4-Flash 在 NPU 上从 ZBAL 迁移到 DEEPEP + war-barrier 架构的具体配置差异和参数选择 (如 mem-fraction-static, kv-cache-dtype)。
 3. 测试执行脚本如何集成硬件供应商的特定环境。

功能与动机

根据 PR body 的描述, 变更动机有二:

1. kimi_k2_6 测试稳定性: 原测试使用 --chunked-prefill-size 262144, 导致完整 64K token 输入被一次性预填充, 峰值激活内存过高, 在 A3 设备上易引发重计算或 OOM, 使夜间延迟测量不稳定。降低至 32768 可稳定预填充, 恢复可复现的端到端延迟。
2. DSV4-Flash测试配置过时: DSV4-Flash的性能准确性测试仍使用旧的ZBAL环境配置。为保持一致性并利用新特性, 需要切换到 DEEPEP + war-barrier 配置, 并配合调整内存分片、预填充 token 数、KV 缓存数据类型、预热请求数等启动参数。同时, 测试执行脚本 run_npu_testcase.sh 需要加载供应商自定义算子环境。

实现拆解

本次变更的核心入口是多个 NPU 测试用例的配置文件, 以及一个测试执行脚本。实现过程可拆解如下:

1. 修复 Kimi K2.6 性能测试预填充稳定性:
 - 涉及文件: test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_16p_in64k_out1k_100ms.py
 - 关键变更: 将启动参数 --chunked-prefill-size 从 262144 修改为 32768。

- 原因与影响：减小预填充块大小以避免单块过大的激活内存，这是修复 A3 设备上测试不稳定性的直接手段，确保性能指标的可复现性。

2. 同步 DSV4-Flash 测试环境配置至 DEEPEP + war-barrier 架构：

- 涉及文件：四个 DSV4-Flash 测试文件（三个性能测试，一个准确性测试）。
- 核心配置替换：在所有测试的环境变量字典中，移除了旧的 ZBAL 相关变量（如 HCCL_BUFFSIZE, SGLANG_ZBAL_*），并统一添加了新的 DEEPEP 和 war-barrier 配置（DEEPEP_HCCL_BUFFSIZE: 2048, SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK: 35, SGLANG_ENABLE_WAR_BARRIER: 1, SGLANG_FORCE_COARSE_WAR_BARRIER: 1）。
- 启动参数调整：
- 内存与计算：将 --mem-fraction-static 从 0.7 或 0.62 调整为 0.68，并添加 --max-prefill-tokens 80000 以控制预填充规模。
- 数据类型与预热：将 --kv-cache-dtype 从 auto 改为 bfloat16；将 warmup_requests 从 0 增加至 16，以稳定首次推理性能。
- 分布式预填充设置：对于 PD 分离部署（1P+1D），将预填充节点的 --chunked-prefill-size 从 -1（不限制）改为 131072，同时将解码节点的 --chunked-prefill-size 从 32768 改为 -1，实现了更精细的控制。并从解码节点的 CUDA Graph batch size 列表中移除了 56。
- 新增环境变量：在 DSv4-Flash 单节点 8p 测试中，添加了 SGLANG_SET_CPU_AFFINITY 和 HCCL_OP_EXPANSION_MODE 变量。在 1P+1D 部署的解码节点环境中，添加了 SGLANG_NPU_USE_MULTI_STREAM。

3. 为 DSV4-Flash 测试引入供应商自定义算子环境：

- 涉及文件：python/sglang/test/ascend/e2e/run_npu_testcase.sh
- 关键变更：在设置完 CANN 和 NNAL 环境后，新增两行 source 命令，加载 /usr/local/Ascend/ascend-toolkit/latest/opp/vendors/customize/bin/set_env.bash 和 ../custom_transformer/bin/set_env.bash，并使用 ll true 容忍可能的文件不存在错误。
- 原因与影响：此变更确保了 DSV4-Flash 测试在 CI 环境中能够找到并链接到其依赖的供应商定制化算子库，是测试能够正确运行的必要前提。

关键文件：

- test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py（模块 DSV4-Flash 测试；类别 test；类型 test-coverage；符号 DEEPSEEK_V4_FLASH_W8A8_8P_ENVS, DEEPSEEK_V4_FLASH_W8A8_8P_OTHER_ARGS）：DSV4-Flash 单节点性能测试的核心配置文件，展示了从 ZBAL 到 DEEPEP + war-barrier 的完整环境变量与启动参数迁移，是本次配置同步的代表性样本。
- test/registered/npu/performance/kimi_k2_6/test_npu_kimi_k2_6_w4a8_16p_in64k_out1k_100ms.py（模块 Kimi K2.6 测试；类别 test；类型 test-coverage；符号 OTHER_ARGS）：Kimi K2.6 性能测试文件，其核心修复点（调整 chunked-prefill-size）是解决测试不稳定性的直接证据。
- python/sglang/test/ascend/e2e/run_npu_testcase.sh（模块 测试执行脚本；类别 test；类型 configuration）：测试执行入口脚本，新增了加载供应商自定义算子环境的命令，是

DSV4-Flash 测试能在 CI 环境中运行的关键配套改动。

- test/registered/npu/accuracy/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_gpqa.py (模块 DSV4-Flash 准确性测试; 类别 test; 类型 test-coverage; 符号 DEEPSEEK_V4_FLASH_W8A8_DSPARK_8P_ENVS) : DSV4-Flash 准确性测试配置文件, 其环境变量的同步迁移与性能测试一致, 确保了准确性测试也在新架构下运行。
- test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in32k_out1k_50ms.py (模块 DSV4-Flash 测试; 类别 test; 类型 test-coverage; 符号 DEEPSEEK_V4_FLASH_W8A8_8P_ENVS, DEEPSEEK_V4_FLASH_W8A8_8P_OTHER_ARGS) : DSV4-Flash 另一个长度 (32k) 的性能测试配置文件, 其变更内容与 in8k 测试高度一致, 体现了配置同步的全面性。
- test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_1p1d_16p_in8k_out1k_50ms.py (模块 DSV4-Flash 分离部署测试; 类别 test; 类型 test-coverage; 符号 DEEPSEEK_V4_FLASH_W8A8_1P1D_PREFILL_ENVS, DEEPSEEK_V4_FLASH_W8A8_1P1D_PREFILL_ARGS, DEEPSEEK_V4_FLASH_W8A8_1P1D_DECODE_ARGS) : DSV4-Flash PD 分离 (1P+1D) 部署的性能测试配置文件, 展示了对预填充和解码节点更精细化的参数调整策略。

关键符号: 未识别

关键源码片段

[test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py](#)

DSV4-Flash 单节点性能测试的核心配置文件, 展示了从 ZBAL 到 DEEPEP + war-barrier 的完整环境变量与启动参数迁移, 是本次配置同步的代表性样本。

变更后的 DSV4-Flash W8A8 单节点 8p 测试环境变量与启动参数配置片段

```
DEEPSEEK_V4_FLASH_W8A8_8P_ENVS = {
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "INF_NAN_MODE_FORCE_DISABLE": "1",
    "SGLANG_SET_CPU_AFFINITY": "1",
    "HCCL_SOCKET_IFNAME": "lo",
    "GLOO_SOCKET_IFNAME": "lo",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "SGLANG_NPU_USE_MULTI_STREAM": "1",
    # deeppep
    "DEEP_NORMAL_MODE_USE_INT8_QUANT": "1",
    "DEEPEP_HCCL_BUFFSIZE": "2048",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "35",
    # war barrier
    "SGLANG_ENABLE_WAR_BARRIER": "1",
    "SGLANG_FORCE_COARSE_WAR_BARRIER": "1",
    # skip gpu branch, mtp 等其他变量 ...
}

DEEPSEEK_V4_FLASH_W8A8_8P_OTHER_ARGS = [
```

```

# ... 前序参数不变 ...
"--prefill-max-requests",
160,
"--max-prefill-tokens",
80000, # 新增：控制预填充的最大 token 数
"--attention-backend",
"dsv4",
# ... 中间参数不变 ...
"--mem-fraction-static",
0.68, # 从 0.7 降低，为 KV 缓存预留更多空间
"--chunked-prefill-size",
131072,
# ... 后续参数不变 ...
"--kv-cache-dtype",
"bfloat16", # 从 "auto" 显式指定为 bfloat16
# ... 其他参数不变 ...
]

```

```

class TestNPUDeepSeekV4FlashW8A88PIn8kOut1k50ms(TestNpuPerformanceTestCaseBase):
    # ... 类属性 ...
    warmup_requests = 16 # 从 0 增加到 16，用于稳定首次推理性能
    # ...

```

python/sglang/test/ascend/e2e/run_npu_testcase.sh

测试执行入口脚本，新增了加载供应商自定义算子环境的命令，是 DSV4-Flash 测试能在 CI 环境中运行的关键配套改动。

```

# 在 run_npu_testcase.sh 中，为适配 DSV4-Flash 自定义算子而新增的环境设置
# set environment of cann
. /usr/local/Ascend/cann/set_env.sh
. /usr/local/Ascend/nnal/atb/set_env.sh
# Adapt Deepseek-V4-Flash test cases with additional environment variables.
source /usr/local/Ascend/ascend-toolkit/latest/opp/vendors/customize/bin/set_env.bash || true
source /usr/local/Ascend/ascend-toolkit/latest/opp/vendors/custom_transformer/bin/set_env.
bash || true

# 后续执行测试用例的逻辑不变 ...

```

评论区精华

本 PR 的 review 仅有批准状态，没有实质性的技术讨论。审核者 [cherryblo](#) 和 [sglang-npu-bot](#) 直接批准了更改。这表明变更可能是在外部沟通（如 Slack）中达成共识，或者变更内容（测试配置同步）的意图明确且风险可控。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 测试稳定性风险（中）：虽然变更旨在提升测试稳定性，但调整多个底层环境变量（如 DEEPEP_HCCL_BUFFSIZE, SGLANG_ENABLE_WAR_BARRIER）和启动参数，如果配置不当，可能导致测试在新配置下出现新的性能波动或失败。需要观察后续 CI 结果。
 2. 配置一致性风险（低）：将多个测试文件的配置从旧架构迁移到新架构，需要确保所有必要的参数都已正确同步。本 PR 看起来已全面覆盖了 DSv4-Flash 的相关测试。
 3. 硬件与驱动兼容性风险（低）：新增的 `source .../vendors/.../set_env.bash` 命令依赖于特定的 Ascend 工具包目录结构。如果 CI 镜像中的路径变更，脚本会因 `|| true` 而静默失败，导致 DSv4-Flash 测试因缺少算子环境而出错。这是一种可预见的维护点。
- 影响：用户与系统影响：
 - 直接影响范围：NPU 平台的 nightly CI 测试流程，特别是 Kimi K2.6 和 DeepSeek V4-Flash 模型的性能与准确性基准测试。
 - 影响程度：间接影响。改善了 CI 的可靠性和测试结果的可比性，确保性能和精度数据反映的是最新硬件与软件栈的真实水平。不直接影响运行时代码或生产服务。团队影响：
 - 为 NPU 平台的性能优化和模型支持工作提供了更可靠的验证基础。同步了多个测试用例的配置，减少了维护不同版本配置的负担。脚本改动可能影响负责 NPU CI 的工程师的调试流程。
 - 风险标记：测试配置同步，CI 基础设施依赖

关联脉络

- PR #37713 [AMD] Fix DSv4 draft extend taking the target compression path during prefill: 同为 DeepSeek V4 模型在特定硬件（AMD）上的 bugfix，与本 PR 在模型支持（DSv4）和硬件平台（NPU vs AMD）上形成关联，反映了社区在多个后端上持续优化该模型的推理稳定性。