PR #2047 完整报告

radixark/miles

fix glm47-flash: use paged MLA prefill on B200

合并时间：2026-08-03 10:38

原文链接：http://prhub.com.cn/radixark/miles/pull/2047

执行摘要

- 一句话：B200 上 GLM-4.7-Flash 改用 paged MLA prefill 修复 rollout
- 推荐动作：值得一读，但主要是启动器层面的硬件适配，不需要深挖核心库。可关注两个设计点：一是用「launcher 参数按硬件分支」而非改核心 backend 来规避硬件兼容问题，保持了核心库的整洁；二是显式保留 sglang_attention_backend 透传口子，给用户留了绕过默认路径的逃生舱。后续建议在 sglang 升级时重点回归 B200 的 MLA prefill 路径。

功能与动机

PR body 明确描述症状：B200 rollout prefill 失败，报错 Unsupported head dimensions: head_dim_qk=256, head_dim_vo=256。根因是 flashinfer_mla_backend.py 在 SM100 上选择了 ragged CUTLASS prefill 路径，而 fmha_cutlass_sm100.cu 不支持 256 维 QK 和 value head。修复目标是让 GLM-4.7-Flash 在 B200 上开箱可用，同时不改变 GPU 拓扑。

实现拆解

1. 扩展参数模型：在 scripts/run_glm47_flash.py 的 ScriptArgs 中将 hardware 从 Literal["H200"] 扩展为 Literal["H200", "B200"]，默认仍为 H200；新增 sglang_attention_backend: str | None = None，允许显式指定 SGLang attention backend。
2. 按硬件差异化 rollout 并行度：sglang_args 中的 --rollout-num-gpus-per-engine 由固定值 4 改为 2 if args.hardware == "B200" else 1。原因是 GLM-4.7-Flash 有 20 个 attention head，rollout TP 必须整除 20；且 B200 上需要 TP2 配合 paged MLA 路径。注意 H200 的默认值从 4 变为 1，这会改变 H200 上的推理吞吐表现。
3. 条件追加 SGLang flag：当 sglang_attention_backend 非 None 且非 default 时透传 --sglang-attention-backend；当 hardware == "B200" 且 backend 为自动选择、default 或 flashinfer 时追加 --sglang-flashinfer-mla-disable-ragged，从而避开 SM100 ragged kernel。显式选择 Triton 等其它 backend 时不会追加该 flag。
4. 文档同步：docs/models/glm/glm4-7-flash.md 更新 launcher 默认值说明，明确默认 hardware=H200、新增 sglang_attention_backend=None，并说明 B200 用 --rollout-num-gpus-per-engine 2、H200 用 1。
5. 测试配套：第三个 commit 删除了一个专用的 launcher 参数测试文件，理由是脚本参数是刻意维护的，单测收益有限；最终只靠双 B200 端到端冒烟验证（跑到首个 optimizer step）

。没有新增自动化测试。

关键文件：

- scripts/run_glm47_flash.py（模块 启动脚本；类别 source；类型 core-logic；符号 ScriptArgs, *execute*）：启动器核心改动所在：扩展 hardware 支持 B200、按硬件设置 rollout TP、条件追加 --sglang-flashinfer-mla-disable-ragged 以绕过 SM100 ragged kernel 限制。
- docs/models/glm/glm4-7-flash.md（模块 模型文档；类别 docs；类型 documentation）：同步说明 launcher 默认值、B200 支持以及 rollout 并行度差异，避免文档与脚本行为脱节。

关键符号：ScriptArgs, *execute*

关键源码片段

scripts/run_glm47_flash.py

启动器核心改动所在：扩展 hardware 支持 B200、按硬件设置 rollout TP、条件追加 --sglang-flashinfer-mla-disable-ragged 以绕过 SM100 ragged kernel 限制。

```
"""GLM-4.7-Flash 启动器：B200 上强制走 paged MLA prefill。"""

from dataclasses import dataclass
from typing import Literal

import miles.utils.external_utils.command_utils as U


@dataclass
class ScriptArgs(U.ExecuteTrainConfig):
    # 只有 H200/B200 两个合法值，默认仍是 H200，保持旧行为。
    hardware: Literal["H200", "B200"] = "H200"
    # 允许用户显式选择 SGLang attention backend，None 表示交给 SGLang 自动选。
    sglang_attention_backend: str | None = None


def execute(args: ScriptArgs):
    # GLM-4.7-Flash 有 20 个 attention head，rollout TP 必须整除 20。
    # B200 上用 TP2 配合 paged MLA prefill；H200 上保持 TP1 以换取吞吐。
    sglang_args = (
        f"--rollout-num-gpus-per-engine {2 if args.hardware == 'B200' else 1} "
        "--sglang-mem-fraction-static 0.7 "
        "--sglang-speculative-algorithm EAGLE "
        "--sglang-speculative-num-steps 2 "
        "--sglang-speculative-eagle-topk 1 "
        "--sglang-speculative-num-draft-tokens 3 "
        "--use-rollout-routing-replay "
    )

    # 显式指定非默认 backend 时透传，方便用户切到 Triton 等替代实现。
```

```
if args.sglang_attention_backend not in (None, "default"):
    sglang_args += f"--sglang-attention-backend {args.sglang_attention_backend} "

# B200 上 FlashInfer 的 SM100 ragged prefill kernel 不支持 head_dim=256 的
# QK 与 VO，只有自动选择 / 默认 /flashinfer 三种情况才需要禁用 ragged 路径；
# 若用户显式选了其它 backend（如 Triton），则无需追加该 flag。
if args.hardware == "B200" and args.sglang_attention_backend in (None, "default",
"flashinfer"):
    sglang_args += "--sglang-flashinfer-mla-disable-ragged "
```

评论区精华

reviewer yueming-yuan 快速 approve，并提出 nit："i think we might not need test for scripts args, given we will only edit it intentionally"。这与提交历史中删除 launcher 参数测试文件的 commit 完全对应，说明团队默认 scripts/ 下的启动器参数属于有意维护的配置，不为它背书自动化测试。

- 是否需要为 scripts 启动器参数编写测试 (testing): 作者采纳该建议，在后续 commit 中删除了专用的 launcher 参数测试文件，只保留双 B200 运行时冒烟验证。

风险与影响

- 风险：
  1. H200 并行度变化：patch 中 --rollout-num-gpus-per-engine 从固定 4 改为 H200=1、B200=2，PR body 声称 "node8/TP4/EP8/rollout4 remain"，但与最终实现并不完全吻合，H200 的 rollout 引擎数实际从 4 降到 1，可能影响 H200 推理吞吐，文档未量化该影响。
  2. 缺少自动化测试：提交中删除了唯一的参数测试，且没有新增针对 --sglang-flashinfer-mla-disable-ragged 的校验，回归只能靠手动冒烟。
  3. 依赖 SGLang 内部 flag：--sglang-flashinfer-mla-disable-ragged 是 SGLang 侧面暴露的开关，后续升级 sglang（本仓库常有此类升级，如 PR#1795）时若 flag 改名或默认行为变化，B200 路径可能再次失效。
  4. 魔法字符串硬件分支：hardware == "B200" 的硬编码在脚本内，未来新增其它 GPU 型号时容易遗漏条件组合。 - 影响：影响范围局限于 scripts/run_glm47_flash.py 启动器和对应文档页。受益方是在 B200 上运行 GLM-4.7-Flash 训练的团队，从「完全跑不起来」变为「可端到端跑到 optimizer step」；H200 用户行为路径保持不变，但 rollout 并行度有变化，需要观察吞吐表现。对 miles 核心库、session、router 等模块零影响，团队协作上也没有引发争议。 - 风险标记：硬件硬编码分支，缺少自动化测试，H200 默认并行度变更，依赖 SGLang 内部 flag

关联脉络

- PR #2040 [AMD] DeepSeek-V4: use the ROCm precision-parity norm path: 同为 scripts 启动器中按硬件（ROCm）分支选择不同实现路径的适配模式，与 2047 的 B200 硬件分支思路一致。

- PR #1733 [AMD] Drop inert DSv4 rollout knobs and add an MTP recipe: 同样是脚本层按硬件 / 算法调整 rollout 配置，说明 miles 的硬件适配主要沉淀在启动脚本层。
- PR #2009 [docs] Add Inkling-Small model page: 与本次 docs 变更同属模型文档维护，反映模型文档与启动脚本同步更新的惯例。