# PR #33130 完整报告

sgl-project/sglang

Disable breakable CUDA graph for NemotronH

合并时间：2026-08-02 04:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33130

## 执行摘要

- 一句话：禁用 NemotronH 的 BCG 规避 prefill 重放精度波动
- 推荐动作：建议快速精读这一段集中式 BCG 兼容性规则表，它展示了 sglang 如何处理模型级捕获缺陷：先切配置期开关止血，再做根因修复。当前应优先与 elvischenv 确认 main 上的修复是否已覆盖 NemotronH 三个架构并补一个回滚 PR，同时补上 GSM8K 波动回归测试，避免再次踩坑。

## 功能与动机

PR 描述明确写到 "Disable breakable CUDA graph (BCG) for NemotronH. This bug needs to be fixed later, still can't find the clear issue"，并附两组 GSM8K 实验：非投机模式三次准确率 0.970 / 0.945 / 0.950，EAGLE 投机模式 0.965 / 0.880 / 0.890 且出现 Invalid 0.005，说明 BCG 重放会带来随机性精度劣化。elvischenv 在 issue 评论中进一步说明 "Seems the issue has been fixed on main. Could we revert this?"，佐证该 PR 是临时开关。

## 实现拆解

1. 架构识别新增：python/sglang/srt/configs/model_config.py 新增 is_nemotron_h(config)，通过 _hf_arch(config) 匹配 NemotronHForCausalLM、NemotronHPuzzleForCausalLM、NemotronHForCausalLMMTP 三个架构名，为后续配置决策提供统一数据契约。
2. 配置期禁用接入：python/sglang/srt/server_args.py 的 _disable_breakable_cudagraph_if_incompatible() 将 is_nemotron_h 与 is_deepseek_dsa、is_deepseek_v4 一并导入，并在规则列表中插入 ("NemotronH (hybrid Mamba2 prefill)", lambda: is_nemotron_h(self.get_model_config().hf_config))；命中任一规则即把 self.cuda_graph_config.prefill.backend 置为 Backend.DISABLED，在 ServerArgs 解析阶段关闭 BCG prefill。
3. 规则归并：新规则与 MLA、DSV4、CP / DCP、TBO、非 DeepEP a2a、多模态等规则共用同一张规则表，删除该条目即可回滚。
4. 测试与配置配套：本 PR 未新增测试、文档或配置项；验证依赖 PR body 中的 GSM8K 手工实验，缺少可自动回归的基线。

关键文件：

- python/sglang/srt/configs/model_config.py（模块 配置解析；类别 source；类型 data-contract；符号 is_nemotron_h）：新增 is_nemotron_h 架构识别函数，作为 BCG 禁用规则的数据契约，是本次变更的基础。
- python/sglang/srt/server_args.py（模块 启动参数；类别 source；类型 dependency-wiring；符号 _disable_breakable_cudagraph_if_incompatible）：在 BCG 兼容性规则表中新增 NemotronH 禁用规则，命中后把 prefill 后端置为 DISABLED，是本次变更的接入点。

关键符号：is_nemotron_h, _disable_breakable_cudagraph_if_incompatible

## 关键源码片段

### python/sglang/srt/configs/model_config.py

新增 is_nemotron_h 架构识别函数，作为 BCG 禁用规则的数据契约，是本次变更的基础。

```
# python/sglang/srt/configs/model_config.py

def is_nemotron_h(config) -> bool:
    # 从 huggingface config 的 architectures 字段读取架构名，
    # 统一识别 NemotronH 家族（基础版、Puzzle 变体、MTP 变体）。
    # 这三个架构共享混合 Mamba2 结构，后续 BCG 兼容性判断
    # 直接以该函数的返回值作为开关依据。
    return _hf_arch(config) in (
        "NemotronHForCausalLM",
        "NemotronHPuzzleForCausalLM",
        "NemotronHForCausalLMMTP",
    )
```

### python/sglang/srt/server_args.py

在 BCG 兼容性规则表中新增 NemotronH 禁用规则，命中后把 prefill 后端置为 DISABLED，是本次变更的接入点。

```
# python/sglang/srt/server_args.py — 摘取规则表核心片段
from sglang.srt.configs.model_config import (
    is_deepseek_dsa,
    is_deepseek_v4,
    is_nemotron_h,
)

rules = [
    # MLA prefill 在 BCG 下走 forward_mha，缺少 eager break；
    # DSA 因为 indexer 本身就按 eager 拆分所以豁免。
    (
        "MLA attention (non-DSA)",
        lambda: self.use_mla_backend()
        and not is_deepseek_dsa(self.get_model_config().hf_config),
    ),
    # NemotronH 的混合 Mamba2 prefill 在 BCG 重放时，状态幕布
```

```
        # 写入没有绑定到捕获缓冲区，会把未写的缓存槽提交出去，
        # 表现为 GSM8K 精度在 0.88~0.97 之间随机波动。先整体禁用。
        (
            "NemotronH (hybrid Mamba2 prefill)",
            lambda: is_nemotron_h(self.get_model_config().hf_config),
        ),
        # DeepSeek-V4 本身兼容 BCG，但 c4 indexer scratch 常驻
        # 捕获池会引发 OOM，也在配置期直接关闭。
        (
            "DeepSeek-V4 (heavy capture-pool memory pressure)",
            lambda: is_deepseek_v4(self.get_model_config().hf_config),
        ),
    ]

    for _name, predicate in rules:
        if predicate():
            self.cuda_graph_config.prefill.backend = Backend.DISABLED
```

## 评论区精华

讨论集中在 PR 合并后的 issue 评论中：

- nvpohanh 评论 "cc @elvischenv"，将问题转给 NemotronH / Mamba 相关维护者跟进。
- elvischenv 随后评论 "@b8zhong Seems the issue has been fixed on main. Could we revert this?"，指出根因已在 main 上修复（大概率是 #34184 对 prefill graph 下 stale track rows 的修复），建议评估回滚该禁用；该反馈发生在合并后，未进入本 PR 决策。
- 代码评审层面 mmangkad 直接 APPROVED（无文字）；gemini-code-assist 机器人的评论仅说明其服务已停止，无实质内容。
- 问题转派给 NemotronH 维护者 (other): 未产生技术结论，仅完成关联维护者的转派。
- main 上已修复，建议回滚 (question): 合并后提出，材料中未见对应回滚落地；建议跟进 #34184 的修复覆盖性后再决定。

## 风险与影响

- 风险：
  1. 性能回退：BCG 被整体禁用后，NemotronH prefill 失去分段捕获优化，且 BCG 构造时强制拒绝 memory-saver 路径，长 prompt / 大批量场景可能出现显存挤占或时延上升；PR 未提供 BCG 开启与关闭的吞吐对比数据。
  2. 影响面偏宽：is_nemotron_h 覆盖三个架构，复现证据仅来自 Ultra 550B A55B 模型，Puzzle / MTP 变体可能被连带禁用而无法细分。
  3. 缺少回归防护：没有测试文件锁定精度波动场景，后续若直接删除规则，CI 无法拦截精度回退。
  4. 维护漂移：规则中的 lambda 访问 self.get_model_config().hf_config，而仓库近期正在推进配置袋（config bags）读取迁移，这段逻辑需随重构同步调整。
- 影响：

1. 用户影响：NemotronH 模型用户启动时会静默禁用 BCG prefill，短 prompt 场景吞吐可能下降，但换来精度稳定。
2. 系统影响：改动集中在配置期判断，不依赖运行时调度，启动开销可忽略；规则表是集中式 if 判断，其他模型可低成本复用。
3. 团队影响：留下了明确的后续挂起项——elvischenv 已请求回滚，需要有人确认 #34184 的修复完整性并决定是否移除该规则。 - 风险标记：静态禁用影响面偏大，根因未定位，缺少回归测试，已出现回滚窗口

## 关联脉络

- PR #34184 Fix stale track rows corrupting conv checkpoints under the prefill graph: 同为 prefill graph 下混合 Mamba 状态写坏缓存的问题，大概率是 elvischenv 所指 main 上的根因修复，是本 PR 应回滚的依据。
- PR #33661 [BCG][5/N] MLA Fully Support: 同一张 BCG 兼容性规则表与 prefill_cuda_graph_runner 的持续演进，展示 BCG 能力逐步收敛和兼容性开关的维护方式。
- PR #34189 [DSV4] Fix silent KV corruption when speculative draft tokens > 4: 同为捕获 / 重放期 buffer 未写导致的静默 KV 损坏修复，可为 NemotronH 根因定位提供相邻参考。