PR #36397 完整报告

sgl-project/sglang

[NVIDIA] Tune custom all reduce v2 for sm_107

合并时间：2026-08-27 06:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36397

执行摘要

- 一句话：为 sm_107 调优自定义 allreduce v2 性能
- 推荐动作：建议阅读该 PR，了解如何针对新 GPU 架构调优通信内核的启发式阈值和块数。值得关注的设计决策包括：使用 CUDA 次要版本区分架构（sm_107 与 sm_100 同主版本），以及通过 mc_blocks 字典管理不同 world size 的 multicast 块数。对于需要支持新架构的开发者，这是一个可作为模板的范例。

功能与动机

PR body 中明确说明：'Improve custom all reduce performance on sm_107'。通过添加针对 sm_107 的初始启发式配置，以提升自定义 All-reduce V2 的性能。此前仅 SM90 和 SM100 有调优配置，sm_107 使用默认保守配置或错误的 SM100 配置。

实现拆解

1. 新增 sm_107 配置生成函数：在 custom_all_reduce_v2.py 中添加 _sm107_configs 函数，返回按 world size 索引的配置字典，包含 graph/eager 的阈值（如 crossover、max size）和块数设置（push/pull/multicast blocks）。与 SM90/SM100 不同，sm_107 的配置使用了较大的 crossover 阈值（如 128MB）和不同的 multicast 块数设置，以适配架构特性。
2. 条件分支调整：在 _get_all_reduce_configs 中解包 cuda_minor，并在 cuda_major == 10 分支内判断 cuda_minor >= 7 时返回 _sm107_configs，否则返回 _sm100_configs。这确保了 sm_107 使用专用配置。
3. 更新文档注释：更新 get_all_reduce_config 的 docstring，说明 SM90、SM100 和 SM107 均已调优。变更仅涉及一个源码文件，无测试配套或配置迁移。

关键文件：

- python/sglang/srt/distributed/device_communicators/configs/custom_all_reduce_v2.py（模块 通信层；类别 source；类型 core-logic；符号 _sm107_configs, config）：核心改动文件，新增 sm_107 架构的调优配置，并调整架构选择逻辑。

关键符号：_sm107_configs, _get_all_reduce_configs

评论区精华

评论者 YAMY1234 指出：'Only affect rubin and safe enough to merge'。说明该 PR 影响范围仅限 Rubin 架构，安全性足够，决策上认为可以直接合并。

- 合并安全性 (question): 确认改动仅影响 sm_107 (Rubin) 架构，不影响既有架构。

风险与影响

- 风险：主要风险集中在 sm_107 配置的准确性，特别是 multicast 块数 num_mc_blocks 的设置：对于 world_size >= 4 时，没有显式的 None 保护，若某个 world_size（如 10、12 等）未在 mc_blocks 字典中，将使用 None，可能导致性能回退到默认或异常。虽然当前测试覆盖了 2-8 和 16，但其他 world size 可能性能不佳。此外，该配置仅基于特定硬件（VR）调优，在其他 sm_107 型号（如 Rubin 变体）上可能不是最优；由于改动仅影响 sm_107 路径，对既有架构（Sm90/SM100）无回归风险。
- 影响：影响范围限定在 sm_107（Rubin）架构的 GPU 上，对使用该架构的用户，自定义 allreduce 性能有显著提升（相对 NCCL 最高 3.26 倍，geomean 2.2%~9.5%）。对 SM90/SM100 无影响，对默认配置的架构无影响。从团队角度，该 PR 为新增架构提供了调优基线，未来可扩展。
- 风险标记：新架构配置覆盖不全，缺少针对其他 world size 的覆盖

关联脉络

- PR #30859 dsa: widen the fp8 k-cache quant kernel's token_id to int64: 涉及 sm_107 相关内核优化，本 PR 的 allreduce 配置可能对 sm_107 上的通信性能有帮助。
- PR #36519 GLM-5.3-Flash cookbook: default Blackwell recipes to FP8 KV + TRT-LLM DSA: 涉及 Blackwell 架构优化，与 sm_107 同属 NVIDIA 新架构调优方向。