PR #7073 完整报告

PaddlePaddle/FastDeploy

[OP] support deepgeem for sm103

合并时间：2026-04-01 21:01

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7073

执行摘要

- 一句话：扩展 DeepGemm 对 SM103 架构的支持，适配新 GPU 硬件。
- 推荐动作：建议快速浏览以了解硬件适配模式，无需精读。值得关注的设计决策：使用 >=100 而非特定版本号来支持未来架构，体现了前瞻性设计；但 review 中关于分支合并的讨论值得思考，可借鉴以简化条件逻辑。对于负责量化或 GPU 优化的工程师，此 PR 展示了如何扩展版本特定功能。

功能与动机

根据 PR 标题和 body，动机是“deepgeem 支持 sm103”，即扩展 DeepGemm（一种高性能 GEMM 实现）对 SM103 架构（NVIDIA GPU 计算能力版本）的支持。PR body 中明确说明“deepgeem 适配”，表明这是硬件适配性改进，没有关联 Issue，但上下文显示近期有多个 GPU 相关优化 PR（如 #7126、#7001、#7094），反映团队持续优化 GPU 计算支持。

实现拆解

实现方案集中在两个量化层文件：

1. block_wise_fp8.py：修改 __init__ 方法中的 deepgemm_scale_ue8m0 条件从 get_sm_version() == 100 改为 get_sm_version() >= 100，扩展支持范围；在 deep_gemm_fp8_gemm_nt 函数中，将版本检查从 == 100 改为 >= 100，并添加断言确保 SM100+ 时 x_scale_tensor.dtype 为 paddle.uint8。
2. fp8_utils.py：修改 load_deep_gemm 和 fused_stack_transpose_quant 函数中的版本检查，从 == 100 改为 >= 100，确保 DeepGemm 加载和量化逻辑适配 SM103+。

关键文件：

- fastdeploy/model_executor/layers/quantization/block_wise_fp8.py（模块 Quantization）：核心变更文件，修改 DeepGemm 的版本检查逻辑和添加断言，直接影响 FP8 量化计算在 SM103+ 的启用。
- fastdeploy/model_executor/layers/quantization/fp8_utils.py（模块 Quantization）：辅助变更文件，调整 DeepGemm 加载和量化函数中的版本条件，确保整体兼容性。

关键符号：init, deep_gemm_fp8_gemm_nt, load_deep_gemm, fused_stack_transpose_quant

评论区精华

review 评论中主要讨论：

1. qingqing01 指出代码与 2.5 分支不一致（引用 PR#7081），但未进一步讨论，可能涉及版本同步问题。
2. zoooo0820 提出两个关键点：一是建议合并 >=100 和 ==100 分支，避免 >100 落入 SM90 分支；二是建议在 fused_moe_deepgemm_backend.py 的 group gemm 处也添加类似检查。这些评论未在 PR 中直接解决，但提交历史显示作者通过多次提交调整代码（如“modify sm version condition”），最终移除了添加的断言，可能简化了实现。决策结论似乎是接受当前修改，zoooo0820 最终批准“LGTM”。未解决疑虑包括分支合并建议和额外文件检查，但鉴于 PR 已合并，可能被视为非阻塞或后续处理。

- 版本分支合并建议 (design): 未在 PR 中实施，但最终批准，可能视为优化建议而非必需。
- 代码一致性检查 (correctness): 未进一步讨论，可能不影响当前合并。
- 额外文件检查建议 (design): 未在 PR 中实施，可能留待后续处理。

风险与影响

- 风险：技术风险较低但需注意：
  1. 兼容性风险：将条件从 ==100 改为 >=100 可能意外启用 DeepGemm 在不支持的架构上，但 SM100+ 通常具有相似特性，风险可控。
  2. 代码一致性风险：qingqing01 提到的与 2.5 分支不一致可能引发跨版本维护问题，需确认分支同步策略。
  3. 测试覆盖不足：Codecov 报告 patch 覆盖率为 60%，2 行缺失覆盖，可能影响变更可靠性。
  4. 未采纳建议风险：zoooo0820 提出的分支合并和额外文件检查未实施，可能导致逻辑冗余或遗漏检查，但当前实现功能完整。
- 影响：影响范围有限：
  1. 用户影响：对使用 SM103+ GPU 的用户，DeepGemm 将自动启用，可能提升 FP8 量化计算性能；对现有用户无负面影响。
  2. 系统影响：仅修改量化层逻辑，不影响核心调度或 API，属于底层优化扩展。
  3. 团队影响：延续 GPU 硬件适配趋势，为未来新架构支持铺平道路；代码变更小，易于维护。
- 风险标记：版本条件扩展，测试覆盖不足，分支同步风险

关联脉络

- PR #7126 [Iluvatar] Fix cuda graph error for tp > 1 in ernie models: 同属 GPU 硬件适配和 bugfix，涉及 CUDA Graph 和版本特定问题，反映团队对多平台支持的持续投入。
- PR #7001 [Feature] Support mtp overlap schedule: 同属 GPU 性能优化 PR，涉及调度和计算优化，展示 FastDeploy 在 GPU 计算领域的演进。

- PR #7094 fix cuda graph capture failure in CI test: 同属 GPU 相关修复，涉及 CUDA 和测试，体现硬件兼容性维护的常见模式。