PR #38193 完整报告

vllm-project/vllm

[XPU] Disable xpu graph by default

合并时间：2026-03-26 16:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38193

执行摘要

- 一句话：禁用 XPU graph 默认启用，添加环境变量控制。
- 推荐动作：PR 变更简单，值得快速 review，关注环境变量添加和默认行为更改，对 XPU 用户重要，建议确保文档更新以通知用户新配置。

功能与动机

根据 PR body，升级到 torch 2.11 后，XPU graph（Intel GPU 上的 CUDAGraph）默认启用，但发现有限制，如需要特定驱动版本且不稳定，因此决定默认禁用，用户可通过环境变量启用。

实现拆解

修改两个文件：在 vllm/envs.py 中添加环境变量 VLLM_XPU_ENABLE_XPU_GRAPH，默认设置为 False；在 vllm/platforms/xpu.py 的 check_and_update_config 方法中添加条件检查，如果环境变量未启用，则禁用 cudagraph_mode，并记录警告日志。

关键文件：

- vllm/envs.py（模块 envs）：添加新环境变量 VLLM_XPU_ENABLE_XPU_GRAPH，定义默认禁用，是配置管理的关键部分。
- vllm/platforms/xpu.py（模块 platforms/xpu）：在配置检查中添加逻辑，根据环境变量禁用 XPU graph，直接影响 XPU 平台的图形执行行为。

关键符号：VLLM_XPU_ENABLE_XPU_GRAPH, check_and_update_config

评论区精华

Review 评论中没有实质讨论，只有自动评论和批准。gemini-code-assist[bot] 描述了变更，指出无反馈；其他审核者如 xinyu-intel 和 bigPYJ1151 直接批准。

- 变更描述 (other): 无讨论，变更被自动接受和批准。

风险与影响

- 风险：风险包括：用户可能不知道新环境变量，导致性能下降；如果驱动版本不匹配，启用 XPU graph 可能不稳定；变更可能影响现有依赖默认启用的工作流。

- 影响：影响所有使用 XPU 平台的用户，需要手动设置环境变量来启用 XPU graph 以获取性能优化；系统稳定性提高，但可能牺牲潜在性能增益；团队需要维护此配置选项。
- 风险标记：默认行为变更，环境变量依赖

关联脉络

- PR #36716 [ROCm]: Update rope+kvcache fusion conditions and disable custom op by default: 类似地禁用默认操作以提高稳定性，涉及平台配置调整。
- PR #38116 Relocate Encoder CUDA graph manager: 涉及 CUDA graph 管理，与本 PR 的 XPU graph 配置相关。
- PR #38076 [Revert] Remove DeepGEMM availability check in DeepseekV32IndexerMetadataBuilder: 调整 CUDA graph 相关逻辑，显示项目中对图形执行稳定性的持续关注。