

PR #53195 完整报告

vllm-project/vllm

[ROCm][DI][CI] Enable WideEP Intranode tests

合并时间: 2026-09-09 06:06

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53195>

执行摘要

- 一句话: 为 AMD 分布式推理 CI 启用 WideEP Intranode 测试并优化超时配置。
- 推荐动作: 本 PR 是一个标准的 CI 配置增强, 值得快速浏览。其核心价值在于补充了 WideEP Intranode 模式的测试覆盖, 这对确保 AMD MI300 平台上高级分布式推理功能的稳定性至关重要。建议关注 models.yaml 中新增的 ep: 配置细节, 特别是 cudagraph_mode 和 compilation-config 的选择是否最优。对于维护 AMD 后端 CI 稳定性的团队成员, 建议精读本 PR。

功能与动机

启用 WideEP Intranode 测试以确保在 AMD MI300 平台上 EP8/DP8 模式的分布式推理功能得到持续验证。同时, 调整超时时间以解决多个 Buildkite 步骤在不同节点上并行运行时可能出现的慢速 NFS 问题, 提升 CI 稳定性。

实现拆解

1. 新增 WideEP Intranode 测试作业: 在 pipeline-disagg.yaml 文件中, 为每个模型 (DeepSeek-V3、DeepSeek-R1-MXFP4、Kimi-K2.5-MXFP4、Kimi-K2.6-MXFP4、MiniMax-M3-MXFP8) 在 1P1D 配置下新增了 WideEP 模式 (WIDE_EP_MODE=1) 的测试作业, 分别测试 proxy 和 vllm-router 两种路由类型。这些作业被归类到新的 mi350 · disagg wide-EP intranode (EP8/DP8, 2 nodes) 分组中。
2. 调整所有测试作业超时时间: 将 pipeline-disagg.yaml 中所有已有测试作业 (包括 1P1D 和 2P2D) 的 timeout_in_minutes 从 120 统一提升至 180。这一调整覆盖了该文件中的所有测试步骤。
3. 更新辅助脚本默认值: 在 run-slurm-disagg-test.sh 中, 将 TIME_LIMIT 的默认值从 02:00:00 更改为 03:00:00, 以与新的 CI 步骤超时对齐。同时, 从每个测试步骤的命令中移除了显式传递 SLURM_TIME_LIMIT=02:00:00 参数, 使其使用脚本新的默认值。
4. 为 WideEP 模式补充模型级配置: 在 models.yaml 中, 为所有模型 (DeepSeek-V3、MiniMax-M3-MXFP8、DeepSeek-R1-MXFP4、Kimi-K2.5-MXFP4、Kimi-K2.6-MXFP4) 新增了 ep: (Expert Parallel) 模式的专用配置, 包括预填充和解码阶段的参数 (如 --gpu-memory-utilization、--enforce-eager、--compilation-config)。同时, 为部分模型调整了环境变量 (如为 DeepSeek-V3 和 DeepSeek-R1-MXFP4 新增 VLLM_ROCM_USE_AITER_FP8BMM: "false", 为 MiniMax-M3-MXFP8 新增 VLLM_KV_CACHE_LAYOUT: "HND")。

5. 清理 SLURM 模板：从 run_xPyD_disagg.slurm 中移除了硬编码的 #SBATCH --time=02:00:00 行，使时间限制完全由运行时参数 TIME_LIMIT 控制。

关键文件：

- .buildkite/amd-disagg/pipeline-disagg.yaml（模块 CI 管道；类别 config；类型 configuration）：主 CI 管道定义文件，包含了所有 WideEP 测试作业的新增以及超时时间的全局调整。
- .buildkite/amd-disagg/models.yaml（模块 模型配置；类别 config；类型 configuration）：模型级配置文件，为所有模型新增了 ep：模式专用的性能参数和环境变量，是 WideEP 测试能够运行的关键。
- .buildkite/amd-disagg/run-slurm-disagg-test.sh（模块 测试脚本；类别 other；类型 core-logic）：测试执行的入口脚本，更新了默认的 SLURM 时间限制和健康检查超时注释。

关键符号：未识别

关键源码片段

.buildkite/amd-disagg/models.yaml

模型级配置文件，为所有模型新增了 ep：模式专用的性能参数和环境变量，是 WideEP 测试能够运行的关键。

```
- model: DeepSeek-V3
env:
  VLLM_ROCM_USE_AITER: "1"
  VLLM_ROCM_USE_AITER_MLA: "1"
  VLLM_ROCM_USE_AITER_MOE: "1"
  VLLM_ROCM_USE_AITER_RMSNORM: "1"
  VLLM_ROCM_USE_AITER_FUSION_SHARED_EXPERTS: "0"
  VLLM_ROCM_USE_AITER_FP8BMM: "false" # 新增环境变量，用于 WideEP
  模式下的特定内核配置。
base_flags: "--trust-remote-code --kv-cache-dtype fp8"
prefill:
  tp: "--gpu-memory-utilization 0.85"
  ep: "--gpu-memory-utilization 0.85 --enforce-eager" # 新增 Expert Parallel
  模式的预填充配置。
decode:
  tp: "--gpu-memory-utilization 0.85"
  ep: '--gpu-memory-utilization 0.75 --compilation-config {"cudagraph_mode":"FULL_DECODE_
  ONLY","custom_ops":["+quant_fp8"]}' # 新增 EP 模式解码配置，使用 FULL_DECODE_ONLY
  模式。
```

评论区精华

PR 的评审过程非常简短且顺利。主要讨论点是作者分享了 DI CI 通过的链接，以及团队成员触发了 CI 运行。评审者 [dllehr-amd](#) 直接给予了批准。没有发现关于实现方案的技术争议或设计权衡讨论。

- 暂无高价值评论线程

风险与影响

- 风险:

1. CI 资源消耗增加: 新增了 10 个 WideEP Intranode 测试作业 (5 个模型 * 2 种路由类型), 将显著增加 AMD DI CI 的总体运行时间和资源占用。
2. 超时设置的普适性: 将所有作业的超时统一提升至 180 分钟, 可能掩盖了某些作业本身的性能退化问题。需要观察后续 CI 稳定性。
3. 模型配置正确性: models.yaml 中新增的 ep: 配置 (如 cudagraph_mode 从 PIECEWISE 改为 FULL_DECODE_ONLY) 需要确保与当前 vLLM 代码库在 AMD 平台上的行为一致, 避免引入新的运行时错误。 - 影响: 本 PR 主要影响 vLLM 项目的 AMD 分布式推理持续集成管道。直接影响: CI 流水线中新增了 WideEP 模式的测试覆盖, 提升了 AMD 平台上 EP/DP 特性的验证能力。间接影响: CI 作业数量和总运行时间增加, 可能影响其他 CI 作业的排队和执行时间。团队影响: 为 AMD 后端维护团队提供了更全面的分布式推理配置测试矩阵, 有助于及早发现 WideEP 模式下的兼容性问题。 - 风险标记: CI 资源消耗增加, 超时配置变更

关联脉络

- 暂无明显关联 PR