

PR #47467 完整报告

vllm-project/vllm

[CPU][Build] Enable oneDNN ITT task collection by default for CPU primitive-level profiling

合并时间: 2026-07-03 12:00

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47467>

执行摘要

- 一句话: 默认开启 oneDNN ITT 任务采集以支持 CPU 性能分析
- 推荐动作: 建议合入。变更简单、风险低, 对 CPU 性能分析有显著价值。值得维护者了解 oneDNN 构建选项及 ITT 接口的使用方式, 便于后续类似的可观测性改进。

功能与动机

当前 oneDNN 在 vLLM CPU 后端中静态编译时, `ONEDNN_ENABLE_ITT_TASKS` 被显式设为 `OFF`, 导致 oneDNN 原语在 VTune 等 ITT 追踪工具中不可见, 合并为匿名的 JIT 代码热点, 极大地增加了 CPU 推理性能分析的难度。启用该选项后, oneDNN 原语执行时可被标记为具名 ITT 任务, 便于精确定位性能瓶颈。

实现拆解

1. 修改 `cmake/cpu_extension.cmake` 文件中的 CMake 变量设置, 将 `ONEDNN_ENABLE_ITT_TASKS` 从 "OFF" 改为 "ON"。
2. 该行位于 oneDNN 配置块内, 其余设置 (如原语类型、JIT 分析等) 保持不变。
3. 仅影响编译时的 oneDNN 库构建, 不修改任何运行时逻辑或源代码。
4. 无需额外测试配置或部署配套, 因为 ITT 任务采集在无追踪器附加时是空操作, 且可通过环境变量 `ONEDNN_ITT_TASK_LEVEL=0` 在运行时完全关闭。

关键文件:

- `cmake/cpu_extension.cmake` (模块 构建脚本; 类别 other; 类型 core-logic): 核心变更文件: 唯一被修改的文件, 将 oneDNN ITT 任务采集选项从 OFF 改为 ON, 直接控制 oneDNN 原语的可观测性。

关键符号: 未识别

关键源码片段

`cmake/cpu_extension.cmake`

核心变更文件: 唯一被修改的文件, 将 oneDNN ITT 任务采集选项从 OFF 改为 ON, 直接控制 oneDNN 原语的可观测性。

```
# 在 CPU 扩展构建中配置 oneDNN
if (ENABLE_X86_ISA OR (ASIMD_FOUND AND NOT APPLE_SILICON_FOUND) OR POWER9_
```

```
FOUND)
# ...
set(ONEDNN_ENABLE_PRIMITIVE "MATMUL;REORDER")
set(ONEDNN_BUILD_GRAPH "OFF")
set(ONEDNN_ENABLE_JIT_PROFILING "ON")
# 启用 oneDNN 的 ITT 任务采集，用于 VTune 等分析工具
set(ONEDNN_ENABLE_ITT_TASKS "ON") # 之前为 OFF
set(ONEDNN_ENABLE_MAX_CPU_ISA "ON")
set(ONEDNN_ENABLE_CPU_ISA_HINTS "ON")
set(ONEDNN_VERBOSE "ON")
# ...
endif()
```

评论区精华

无实质性讨论；审核者 bigPYJ1151 已批准，bot 评论为自动化提示。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。这是一个构建选项切换，仅影响 CPU 后端 oneDNN 库的编译。运行时若无 ITT 收集器附加，标记操作为空操作，无性能开销。可通过环境变量 `ONEDNN_ITT_TASK_LEVEL=0` 在运行时完全关闭。不会影响其他平台（如 GPU、XPU）或非 CPU 后端。
- 影响：影响范围限于 vLLM CPU 后端编译产物的可观测性。对于需要进行 CPU 推理性能分析的开发者和用户，VTune 等工具能显示 oneDNN 原语的详细硬件事件数据。对于生产环境用户，无影响，因为 ITT 任务默认仅在有追踪器时激活。
- 风险标记：仅限 CPU 构建，对生产无影响

关联脉络

- PR #47447 [Bugfix][CPU] Ship examples/ in the CPU release image: 同为 CPU 相关构建改进，但内容不同，无直接依赖。