PR #40037 完整报告

vllm-project/vllm

[ROCm] Add gfx1102/gfx1103 support

合并时间：2026-04-23 16:32

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40037

执行摘要

此 PR 为 AMD RDNA 3 架构的 gfx1102 和 gfx1103 GPU（如 Radeon 780M iGPU）添加支持，修复了因架构宏缺失导致的运行时崩溃。核心改动是用编译器提供的通用宏 __GFX11__/__GFX12__ 替代手动枚举单个架构，避免未来遗漏。

功能与动机

gfx1103（RDNA 3，如 Radeon 780M iGPU）未包含在 HIP_SUPPORTED_ARCHS 列表及编译时架构宏中，导致 wvSplitK 内核在 gfx1103 上编译为 UNREACHABLE_CODE（assert false），运行时崩溃。

实现拆解

1. CMakeLists.txt：在 HIP_SUPPORTED_ARCHS 列表中添加 gfx1102 和 gfx1103，确保 CMake 构建系统能识别并编译这些架构的代码。
2. csrc/rocm/skinny_gemms.cu：移除手动枚举各 gfx11xx/gfx12xx 的宏定义，改用编译器提供的 __GFX11__ 和 __GFX12__ 宏（这些宏在对应架构下自动定义）。同时，内核条件编译中的 __HIP__GFX12__ 也替换为 __GFX12__，保持一致性。

```
// Before: manual enumeration of each arch
// #if defined(__HIPCC__) && (defined(__gfx1100__) || defined(__gfx1101__) || ...)
// #define __HIP__GFX1X__
// #endif

// After: use compiler-provided macros that cover all family members
#if defined(__GFX11__) || defined(__GFX12__)
#define __HIP__GFX1X__
#endif

// Kernel condition also updated
#if defined(__HIP__MI3XX__) || defined(__GFX12__)
// ... actual FP8 kernel ...
#else
// ... UNREACHABLE_CODE fallback ...
#endif
```

1. csrc/rocm/attention.cu：同样移除手动定义的 __HIP__GFX11__ 和 __HIP__GFX12__ 宏，在引用处直接使用 __GFX11__ 和 __GFX12__ 宏。

```
// Remove custom macro definitions
// #if defined(__HIPCC__) && (defined(__gfx1100__) || defined(__gfx1101__) || ...)
// #define __HIP__GFX11__
// #endif

// Use compiler macros directly
#elif defined(__GFX11__)
// ... gfx11 attention path ...
#elif defined(__GFX12__)
// ... gfx12 attention path ...
```

评论区精华

无实质讨论，仅有一条自动机器人评论和一条 approve。

风险与影响

风险：低风险。改动为宏替换和列表扩展，不涉及逻辑变更。但需确保所有使用旧宏 __HIP__GFX11__/__HIP__GFX12__ 的地方已全部替换，避免遗漏。本 PR 已覆盖 skinny_gemms.cu 和 attention.cu 两个主要文件。

影响：

- 正向：支持 gfx1102/gfx1103（如 Radeon 780M iGPU），扩大 AMD ROCm 硬件兼容性。
- 无影响：现有 gfx1100/1101/1150 等架构不受影响，NVIDIA/Intel 平台不受影响。

关联脉络

本 PR 与近期 [ROCm] 相关的 PR（如 #39789，XPU 平台类似问题）共同体现了 vLLM 对不同硬件平台的持续适配。采用编译器提供的通用宏是一种更健壮的做法，未来的 gfx11xx/gfx12xx 变体将自动获得支持，无需再次修改源代码。