



ROCM 在 RISC-V 上的 适配情况

报告人：陈璇

日期：2025.09.23



自我介绍

陈璇

- PLCT 实验室工程师
- LibreOffice / Eclipse riscv64 porter

Nickname: @Sakura286

Email: sakura286@outlook.com



目录

-  1. ROCm 支持情况概述
-  2. 用户空间软件栈构建与移植
-  3. 内核部分

第1部分

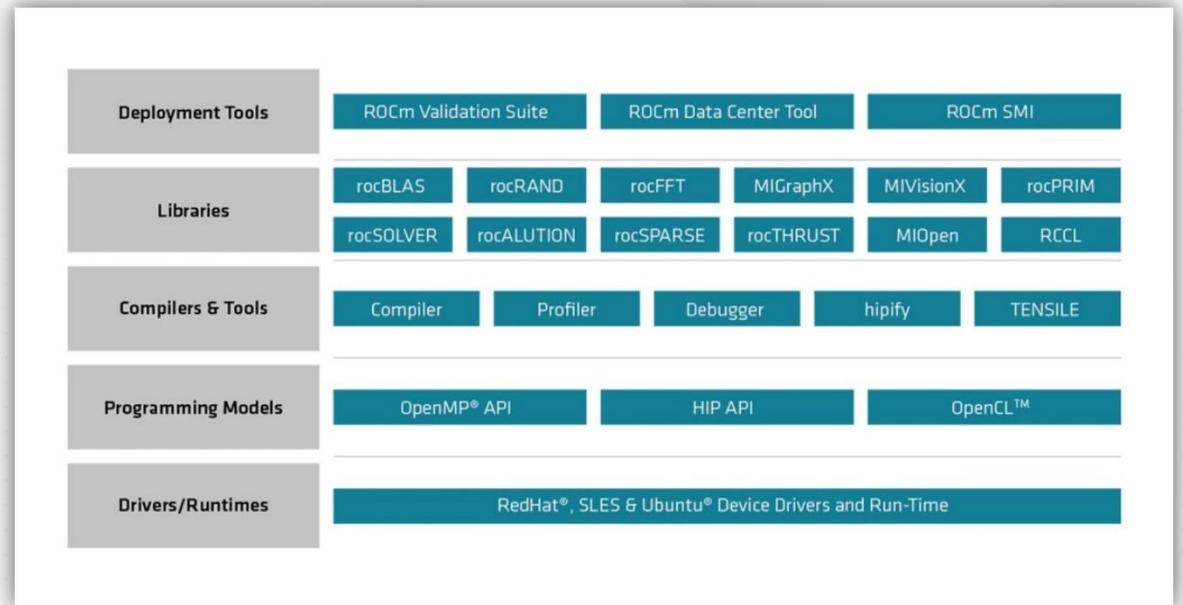
ROCm 支持情况概述



ROCM 简述

ROCM (Radeon Open Compute Platform) 是一个开放式软件栈，包含多种驱动程序、开发工具和 API，可为从底层内核到最终用户应用的 GPU 编程提供助力。ROCM 已针对生成式 AI 和 HPC 应用进行了优化，而且能够轻松将现有代码迁移到 ROCm 软件。

简要来说，ROCM 是 AMD 提供的 GPU 计算平台，类比/对标的是 NVIDIA 的 CUDA。而且 ROCM 是开源软件栈，用户有适配至不同 SOC/OS 的可能。



// 支持情况

GPU支持情况

- RDNA 架构的中高端的 Radeon (Pro) 显卡，及CDNA系列的 AMD Instinct 显卡 [1]

CPU支持情况

- 根据 AMD 官方提供的构建脚本，官方支持的指令集架构为 **x86** [2]
- Debian ROCm 支持的指令集架构为 amd64 arm64 与 ppc64el [3]
- 理论上，支持 **PCIe Atoms** 的现代 CPU 都可以运行

[1] <https://rocm.docs.amd.com/projects/install-on-linux/en/latest/reference/system-requirements.html>

[2] https://github.com/ROCM/ROCM/blob/05a66f75fea71fe19ba29f694c7c22854187e334/tools/rocm-build/build_lightning.sh#L375

[3] <https://buildd.debian.org/status/package.php?p=rocblas&suite=sid>

RISC-V 移植情况简介

llama.cpp

ROCM Libraries

ROCM System

RevyOS(vendor kernel 6.16)

SG2044 + W7800

目前，PLCT 实验室的 RevyOS 小队在 SG2044 上，使用 Radeon Pro W7800 48GB 显卡成功运行**基于 ROCm 6.4.3 后端的 llama.cpp** [1]

RISC-V 成为继 X86、ARM64、PPC64 之后下一个能够搭配 AMD 显卡通过 ROCm 后端运行**异构计算**的 CPU 架构。

[1] <https://zhuanlan.zhihu.com/p/1928513776661038179>

RISC-V 移植情况简介

card	model	size	params	test	t/s
W7800 48GB	llama 70B Q4_K - Medium	39.59 GiB	70.55 B	pp512	236.55 ± 0.22
W7800 48GB	llama 70B Q4_K - Medium	39.59 GiB	70.55 B	tg128	9.71 ± 0.11
W7800 48GB	llama 8B Q4_K - Medium	4.58 GiB	8.03 B	pp512	2043.44 ± 5.71
W7800 48GB	llama 8B Q4_K - Medium	4.58 GiB	8.03 B	tg128	63.72 ± 0.10
W7800 48GB	qwen2 1.5B Q4_K - Medium	1.04 GiB	1.78 B	pp512	7300.93 ± 61.38
W7800 48GB	qwen2 1.5B Q4_K - Medium	1.04 GiB	1.78 B	tg128	94.32 ± 0.06

在 llama.cpp 不断优化 vulkan 后端的前提下，rocm 后端的表现依然要优于 vulkan 大约 50%~100% 的速度

第2部分

用户空间软件栈构建与移植



// 单软件包视角1：配置

添加 RISC-V 的入口

RISC-V 目前并非官方支持架构，所以要在检测 arch 时，仿照其他架构添加相应的编译入口（修改 Makefile 等配置文件）

见 *rocm-llvm*、*rocm-hip-on-rocclr* 仓库

// 单软件包视角2：编译

指令/调用替换

fence.tso -> fence.rw.rw

写屏障：`_mm_sfence(); -> asm volatile("fence w,w" ::: "memory"); [1]`

全内存屏障：`_mm_mfence(); -> asm volatile("fence rw,rw" ::: "memory"); [2]`

线程暂停：`_mm_pause(); -> asm volatile(".insn 0x0100000f" ::: "memory"); [3]`

V 拓展的一点小坑

如果目标平台不支持 V，则编译 llama.cpp 时需要 `-DGGML_RVV=OFF` 来关闭以避免运行时报错 [4]

[1] <https://elixir.bootlin.com/linux/v6.12.6/source/tools/arch/riscv/include/asm/barrier.h>

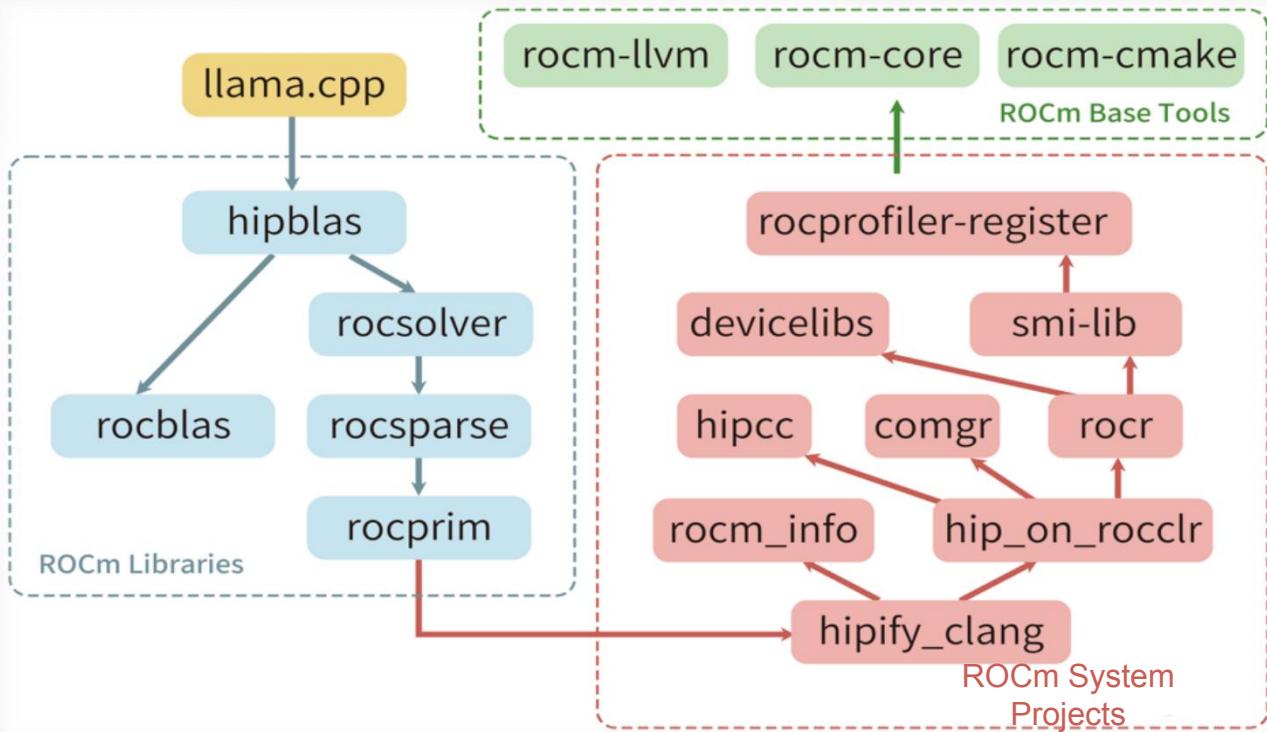
<https://elixir.bootlin.com/linux/v6.12.6/source/arch/x86/um/asm/barrier.h>

[2] RISC-V Memory Consistency Model Specification (DRAFT) - 2.6 Code Porting Guidelines

[3] <https://gcc.gnu.org/pipermail/gcc-patches/2023-August/627546.html>

[4] <https://github.com/ggml-org/llama.cpp/blob/86587da03bd78df8f4e7d8b111a0c1d2494d6ed0/ggml/CMakeLists.txt#L131>

// 多软件包视角1：依赖分析



大致的软件依赖关系可以查看 ROCm Github 仓库[1]，更详细的依赖情况，请查看 revyos-rocm github 组织下仓库[2]的 debian/control

[1] <https://github.com/ROCM/ROCM/blob/develop/tools/rocm-build/ROCM.mk>

[2] <https://github.com/orgs/revyos-rocm/repositories>

// 多软件包视角2：路径

官方仓库的默认产物路径不符合发行版行为。

`opt/rocm` -> `/usr` or `/opt/rocm-${ROCM_VERSION}`

`lib/llvm/bin` -> `bin`

`amdclang` -> `clang`

见*rocm-tensile*、*rocm-rocblas* 仓库

// 多软件包视角3：编译时间控制

编译参数

-DGPU_TARGETS="gfx908:xnack-;gfx90a:xnack-;gfx90a:xnack+;gfx940;gfx941;gfx942;gfx1030;gfx1100"

在编译数学库及 rccl 等库时，这是 hipcc 生成 GPU 代码的默认目标，但因为 Linux 上 AMD 独显并不支持 xnack^[1]，所以 xnack 部分可以去掉。

减少 GPU 代码生成目标可以显著减少编译时间以及产物大小，尤其是在编译 test 及 benchmark 的情况下。

例如，对于 7900XTX 来说，仅需 **-DGPU_TARGETS="gfx1100"**。

-DBUILD_CLIENTS_TESTS=OFF

-DBUILD_CLIENTS_BENCHMARKS=OFF

视情况关闭测试与 Benchmark

[1] <https://niconiconi.neocities.org/tech-notes/xnack-on-amd-gpus/>

第3部分

内核部分



KFD 驱动

ROCM 需要支持 HSA 特性的 amdkfd 驱动支持[1]，虽然 Linux 内核直到 6.16 才正式启用 riscv64 的 amdkfd 编译配置[3]，但 amdkfd 在 riscv64 上的支持早已完善，如有需要可以手动开启

```
config HSA_AMD
    bool "HSA kernel driver for AMD GPU devices"
- depends on DRM_AMDGPU && (X86_64 || ARM64 || PPC64)
+ depends on DRM_AMDGPU && (X86_64 || ARM64 || PPC64 || RISCV)
    select HMM_MIRROR
    select MMU_NOTIFIER
-----
```

[1] HSA, Heterogeneous System Architecture, 异构计算系统

KFD, Kernel Fusion Driver, 内核融合驱动

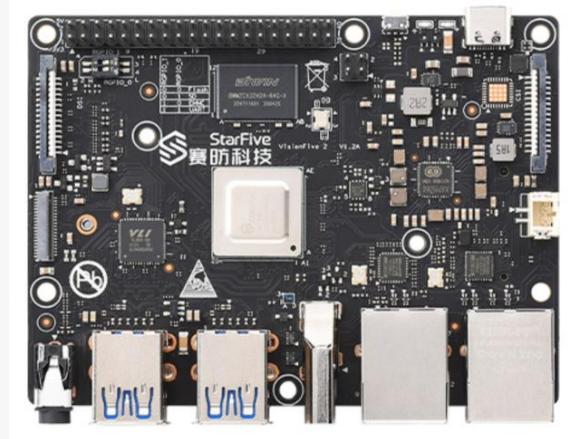
[2] <https://www.phoronix.com/news/AMDKFD-RISC-V-Linux-6.16>

PCIe 一致性

异构计算涉及到 CPU 与显卡间数据的相互通信，需要设备在硬件层面能够保证 PCIe 一致性。目前支持 PCIe 一致性的 RISC-V 设备(SOC)较少，仅有如下几款：



FU740
(Hifive Unmatched)



JH7110
(StarFive VisionFive 2)



SG2044



DP1000
(Milk-V Titan)

用户态软件栈部分

源码

<https://github.com/orgs/revyos-rocm/repositories>

- 构建脚本为 debian/rules 文件
- 构建补丁在 debian/patches 目录下

二进制软件仓库

<https://fast-mirror.isrc.ac.cn/revyos/trixie/revyos-rocm/>

内核部分

RevyOS SG2044 6.1x 内核源码

<https://github.com/revyos/sg2044-vendor-kernel>

mmap() 补丁

https://lore.kernel.org/all/20250707193411886Kc-TWknP0PER2_sEg-byb@zte.com.cn/

感谢观看

本 ppt 获取地址：

<https://github.com/Sakura286/TmpShare/blob/main/20250923-120000-ROCM.pdf>

