ISCAS
中国科学院软件研究所
Institute of Software Chinese Academy of Sciences


# ROCm 在 RISC-V 上的适配情况

## 用户空间软件构建

报告人：陈璇
邮箱：sakura286@outlook.com
日期：2025.09.10

创新求实 · 永竞一流

地址：北京市海淀区中关村南四街4号中国科学院软件园区

# 目录

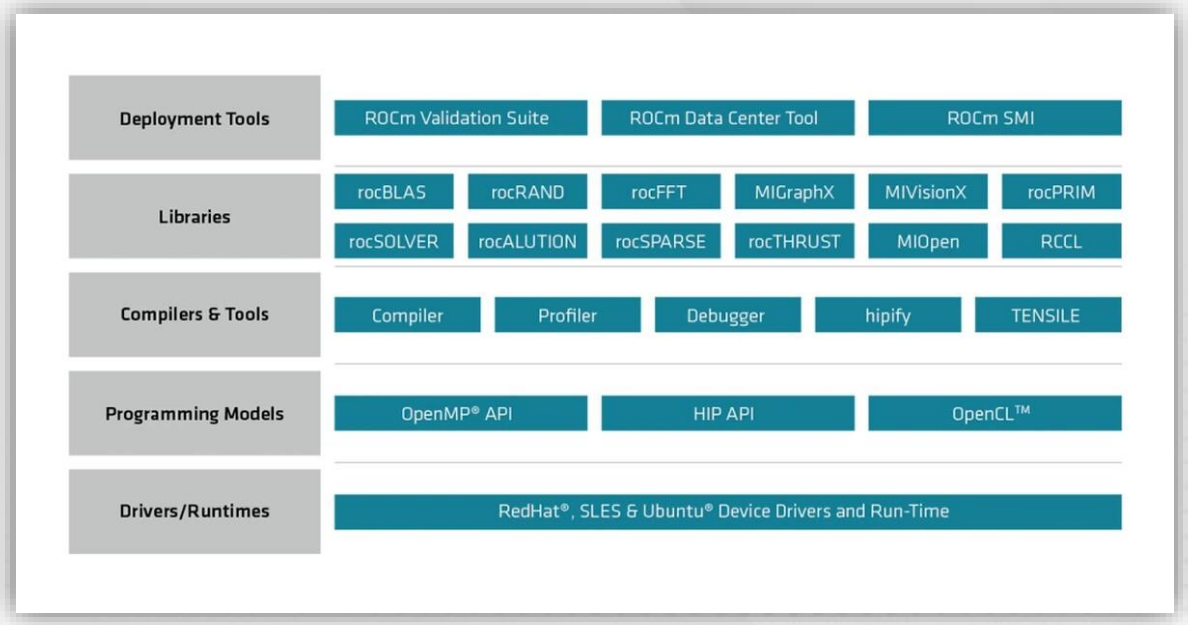
第1部分

# ROCm 支持情况概述

# ROCm 简述

ROCm（**R**adeon **O**pen **C**ompute Platfor**m**）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。



## GPU支持情况

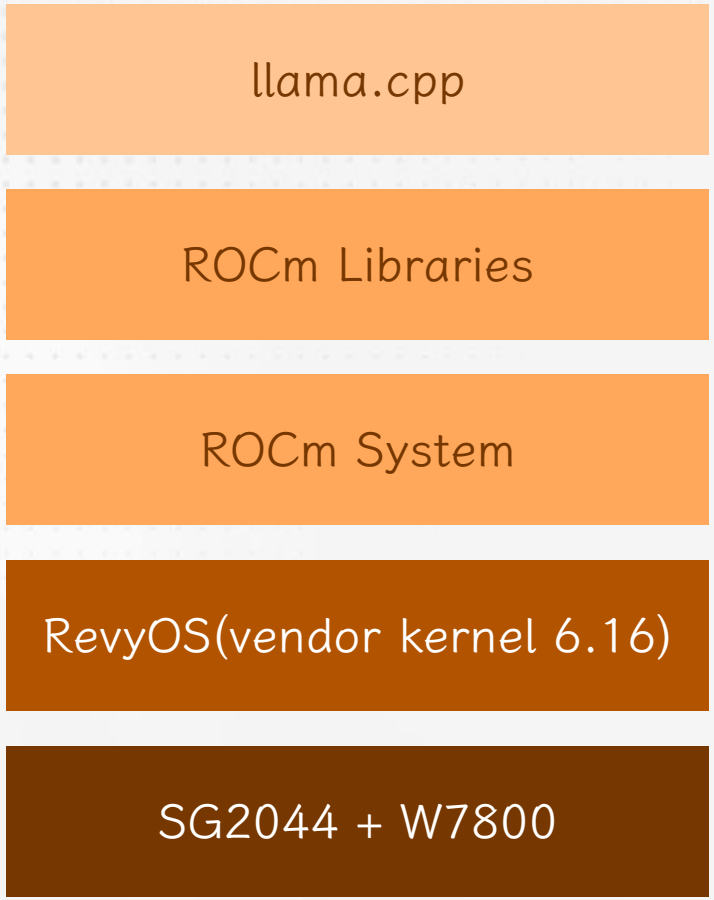
- RDNA 架构的中高端的 Radeon (Pro) 显卡，及CDNA系列的 AMD Instinct 显卡 [1]

## 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




目前，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]

[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 |

第2部分

# 用户空间软件栈构建与移植

# 构建&移植 1：编译入口、参数、路径修复等杂项

## 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入口

*见 rocm-llvm、rocm-hip-on-rocclr 仓库*

## 路径修复

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*

## 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"`

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

# 构建&移植 2：指令/函数替换 & V 拓展

## 指令替换

```
fence.tso -> fence.rw.rw
_mm_sfence(); -> asm volatile("fence w,w" ::: "memory"); [1]
_mm_mfence(); -> asm volatile("fence rw,rw" ::: "memory"); [2]
_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]

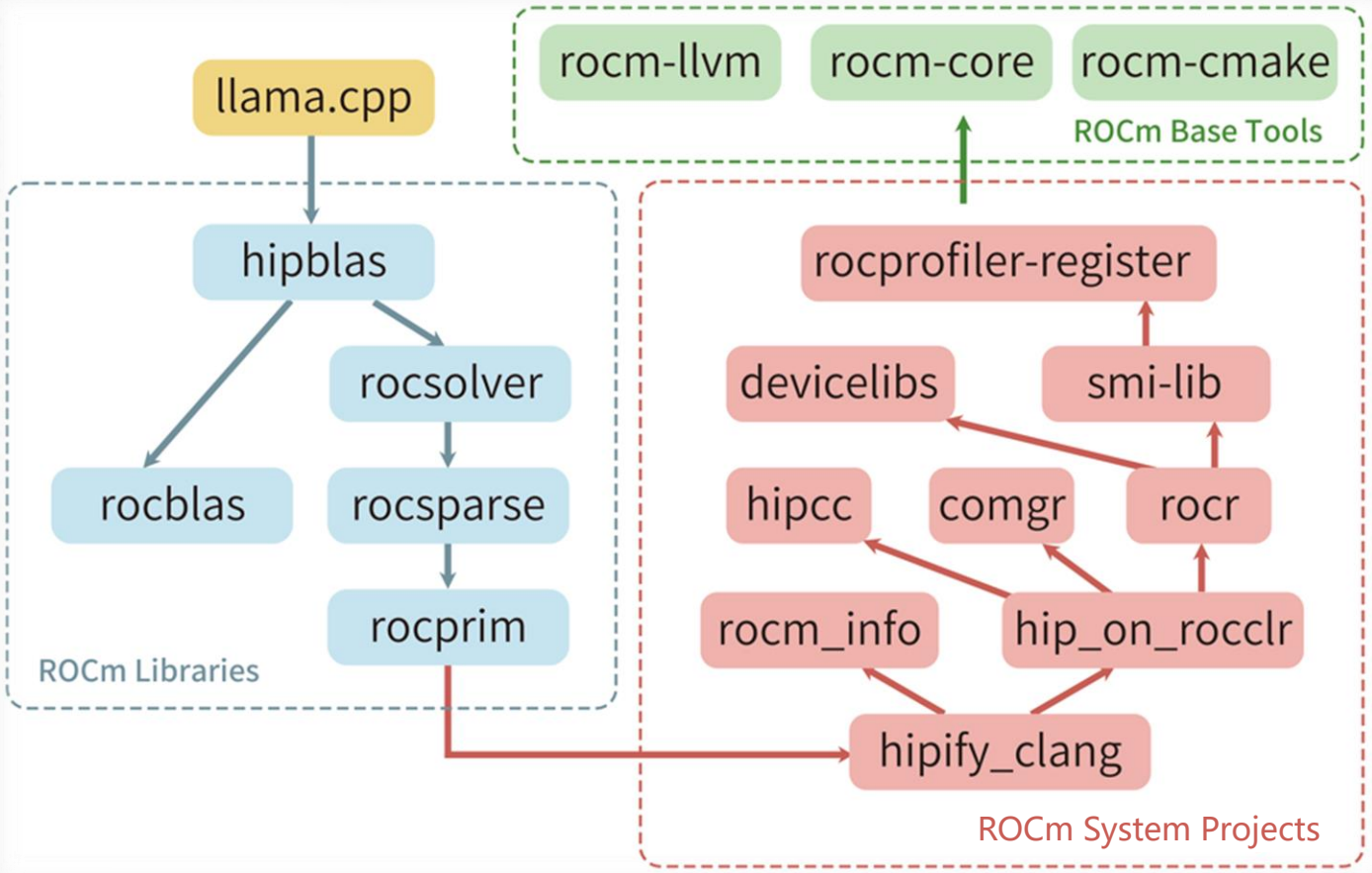

[1] https://elixir.bootlin.com/linux/v6.12.6/source/tools/arch/riscv/include/asm/barrier.h
https://elixir.bootlin.com/linux/v6.12.6/source/arch/x86/um/asm/barrier.h

[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

# 构建&移植 3：llama.cpp 依赖链条




更详细的依赖情况，请查看 revyos-rocm github 组织下仓库的 debian/control

# 源码与产物

## 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

## 二进制软件包仓库

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

第3部分

# 内核部分简要概括

# 源码

RevyOS SG2044 6.16 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/

ISCAS
中国科学院软件研究所
Institute of Software Chinese Academy of Sciences

# 感谢观看

本 ppt 获取地址：

https://github.com/Sakura286/TmpShare/20250910-120000-ROCm-user.pdf