CANN开源社区介绍

大模型专家 冯源
2025年12月20日

content

目录

CANN社区是什么？

CANN是什么？

CANN（Compute Architecture for Neural Networks）是AI异构计算架构，对上支持多种AI框架，对下服务AI处理器与编程，发挥承上启下的关键作用，是提升华为AI处理器计算效率的关键平台

CANN社区

CANN 社区是围绕CANN构建的开源协作平台，提供环境部署指导、开源代码获取、协作开发、技术问答、社区互动、赋能培训等服务，促进成员协作

CANN开源愿景

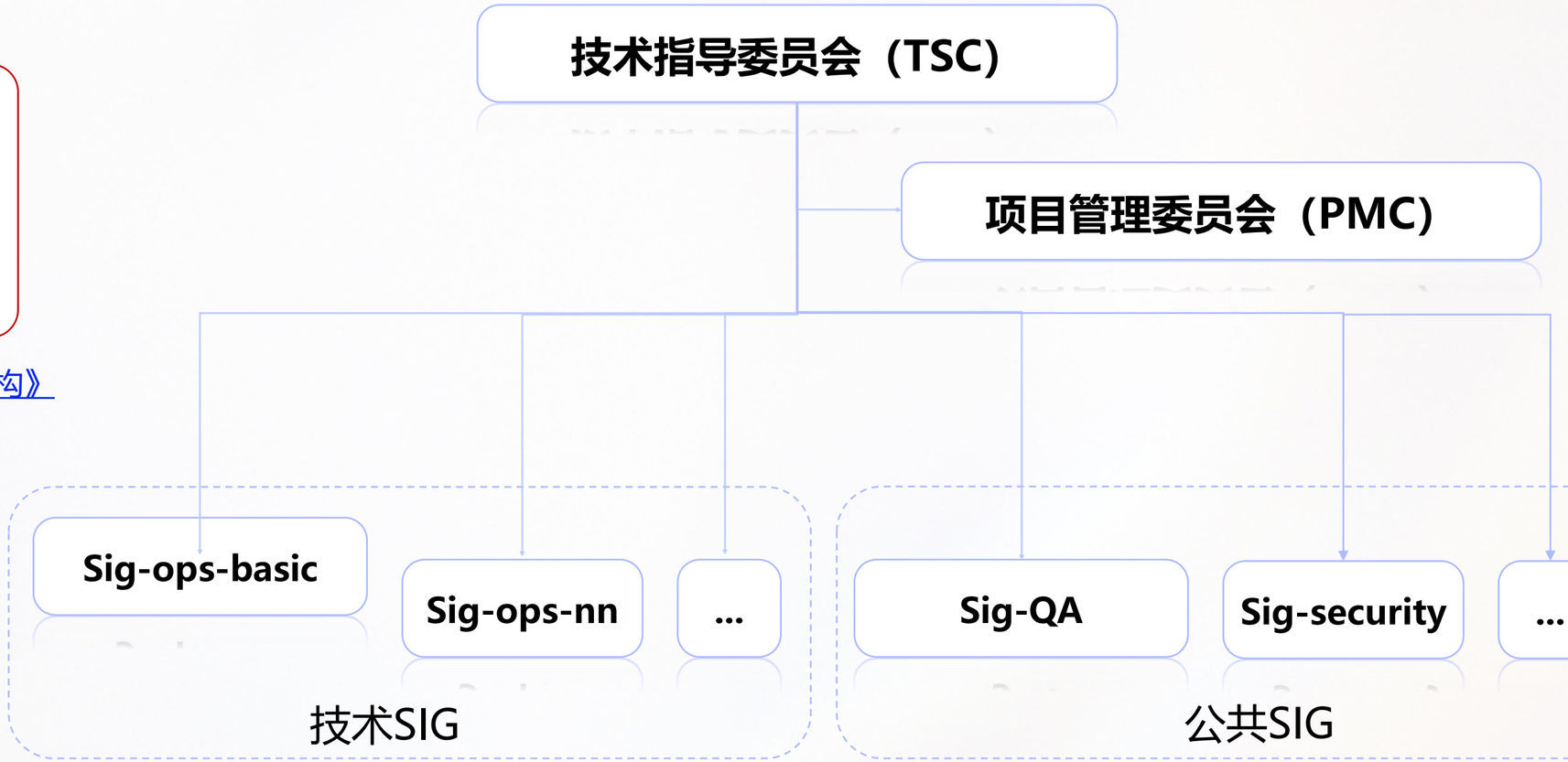
打造开放易用、技术领先的AI算力新生态，成为国内开发者首选的AI开发平台

CANN开源使命

使能开发者基于CANN社区自主研究创新，构筑根深叶茂、跨产业协同共享共赢的AI生态

CANN治理架构

《CANN开源组织架构》



《CANN开源治理制度》

CANN技术架构：打造极致性能、极简易用的AI算力使能层，释放昇腾澎湃算力

❶ 使能大模型并行计算加速
提供高性能算子及通信算法，释放澎湃算力

❷ 高效开发与生态迁移
提供多种算子开发，使能高效开发

❸ 开源开放，生态兼容
提供丰富参考实践，使能自主创新

AI框架

全面支持业界AI框架，适配PyTorch社区版本

PyTorch | 昇思 MindSpore | TensorFlow | 飞桨 | Caffe | Jittor 计图 | ……

CANN 异构计算架构

算子库	通信库	图引擎	领域加速库	工具
大模型融合算子 NN/CV/Math基础算子	集合通信算法 分布式通信	图编译优化 图执行加速	覆盖不同开发场景加速套件	支持算子调试，性能调优，提供可视化能力

算子编程 Ascend C | pyPTO

BiSheng Compiler 毕昇编译器 虚拟指令PTO instruction | 异构编译优化 | AscendNPU IR

Runtime 运行时 控制流 | 内存管理 | 任务调度

Driver 驱动 设备管理 | 加速器驱动 | 板级驱动

昇腾硬件

Ascend 昇腾系列AI处理器 ……

CANN开源开放、分层解耦，满足各层级灵活开发需求

从模型、算子、内核、底层资源等多层级优化和开发，兼顾性能与开发易用性

图模式开发

模型整图下发，降低 Host 调度开销，提升整图执行性能

单算子API调用

框架直调领域加速库或算子库，平滑迁移、高效开发

自定义算子开发

提供 C、C++、Python 等编程方式，匹配不同开发习惯

直调底层Runtime接口

细粒度控制硬件资源，释放硬件性能，支持极致创新

CANN全面开源开放规划

2025

- 解耦并开源算子库
- 开源CATLASS模版库
- 开放AscendNPU IR支持，支持Triton
- 1230 910B/910C 全面开源开放

2026

- 950系列上市即开源
- Ascend C使能下一代处理器950编程特性
- 支持多代际昇腾产品开发和创新

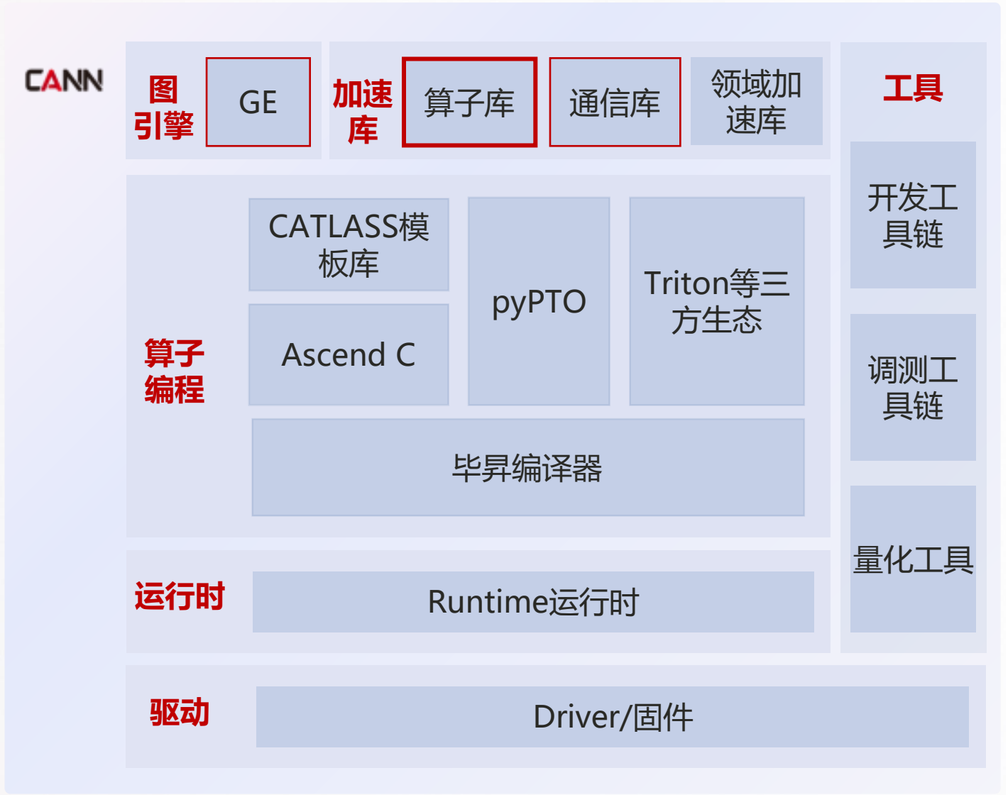
2027

持续迭代期：

未来每代际产品配套软件持续迭代

https://gitcode.com/cann

CANN开源进展：已开源全量算子库、集合通信、编程语言及图引擎



技术领域	主要代码仓	对客户的帮助
算子	ops-transformer	融合算子库，将多个独立的“小算子”融合成一个“大算子”，常用于加速大模型，典型的例子如FlashAttention、以及各种计算通信融合算子
	ops-nn	加速神经网络计算的高阶算子库，涵盖常见的张量matmul、activation、loss计算等操作
	ops-math	提供数学类基础计算的加速，包括math类、conversion类等算子
	ops-cv	图像处理、目标检测等高阶算子库，涵盖常见的图像处理操作，包括image类、objdetect类
	opbase	提供算子公共能力的基础框架库，涵盖aclnn基础框架和公共依赖项
集合通信	hccl	集合通信库，用户可以参考和实现自有集合通信算子/算法
	hcomm	集合通信控制面&数据面，用户可以自行修改通信框架和通信机制，进行维测增强
	hixl	灵活、高效的昇腾单边通信库，面向集群场景提供简单、可靠、高效的点对点数据传输能力
GE图引擎	ge	图引擎，1、图模式实现参考 2、增强开放能力，供用户定制图编译行为
	graph-autofusion	面向昇腾（Ascend）芯片的轻量级、解耦式组件集合，旨在通过自动融合技术加速模型执行。 目前已开源 SuperKernel 组件，未来将持续开放更多自动融合相关模块
	metadef	cann 算子以及图引擎相关的元数据定义，即相关数据结构以及对外接口定义
Ascend C 编程	asc-devkit	Ascend C API和模板库，用户可以自行修改API和模板库的实现，按需封装，提高开发效率。
	asc-tools	Ascend C开发工具，用户可以自行修改和扩展相关工具
	pyasc	Ascend C python前端，支持用户扩展python编程API和优化能力
工具	oam-tools	提供支持典型维测问题的辅助定位工具，包括一键收集npu维测信息、aic error辅助分析和集合通信性能/正确性测试
运行时	npu-runtime	运行时/DFx采集能力，并支持acl Graph图捕获和重放，用户可以自主开展维测，探索运行时和资源管理创新
驱动	driver	HAL/OS适配/设备管理/资源管理等host侧驱动，支撑客户自主创新

930开源开放：支持算子分包独立构建、独立安装升级，提升开发者体验

930社区尝鲜版：支持子包独立安装、独立升级，295个算子完成开源

历史现状

整包模式

930开源

独立升级、独立构建模式

295个↑

开源算子数

50%↓

构建时长优化

85%↓

升级包按需部署

开源试运营完成首个社区外部贡献 & 社区开发者联创case上线

AsNumpy: CANN社区首个完全由社区贡献者开发的代码仓

哈工大苏统华教授团队 ✚ 业界Numpy https://numpy.org/

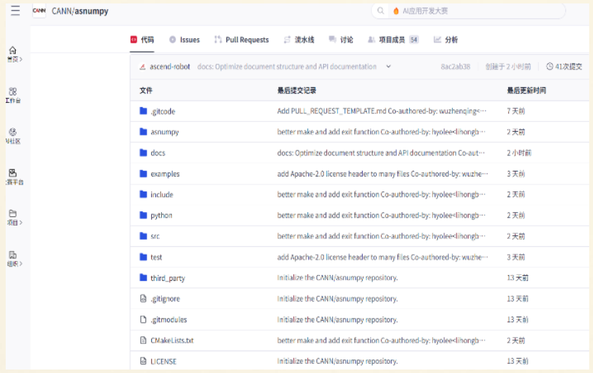
CANN 社区

项目介绍

哈尔滨工业大学计算学部苏统华、王甜甜老师团队联合华为CANN团队开发的华为昇腾NPU原生Numpy仓库

https://gitcode.com/cann/asnumpy

Apache-2.0 C++ 41 提交数

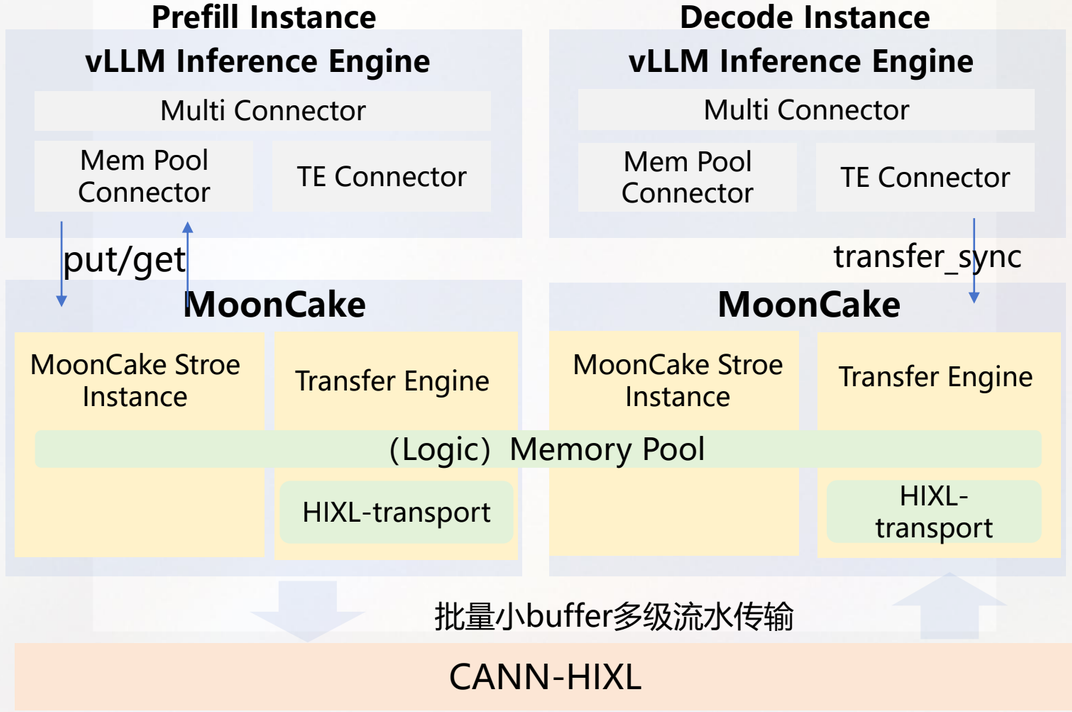


CANN-HIXL: 联合社区开发者共建

三方联合：社区开发者 + CANN-HIXL协同完成TTFT优化40%，并反哺相关优化至社区

- 与社区基于HIXL完成NPU对接
- 协同HIXL完成批量小Buffer多级流水传输方案在CANN的落地
- 贡献昇腾亲和的BatchPut/BatchGet接口至开发者社区

[与 CANN 联合共创 HIXL 组件：开放昇腾底层高速互联，提供简易 API]



0day 支持DeepSeek-V3.2-Exp / Kimi-K2-Thinking模型

[2025/09] CANN社区0day支持昇腾推理部署DeepSeek-V3.2-Exp

- **低比特量化深度优化**：支持 W8A8C8 量化格式，显存占用降低 50% 且精度损失 < 1%；
- **长序列稀疏计算加速**：适配 DSA 稀疏注意力机制，64卡128K 长序列推理 TTFT<2 秒，TPOT<20ms, 吞吐量提升 3 倍；
- **算子融合与硬件适配**：基于 AscendC 实现 LI+SFA 融合 Kernel，释放稀疏计算潜力，配套技术文档与代码已开源；
- **自研PyPTO框架**：依托 PyPTO 框架实现 NPU DSA，提升融合算子编程易用性并扩展 Decode Attention 融合，文档与代码同步开源；

【2025.11】CANN社区0day支持Kimi-K2-Thinking，支持256K长序列

- **Flash Decode加速**：针对小 batch、长序列负载降低时延、提升算力利用率
- **INT4 量化适配**：完成 A16W4（pergroup=32）量化格式适配，配套 GMM 算子开源，平衡速度与精度
- **分布式传输优化**：HIXL 组件开源，与Mooncake社区全面适配，支持多种底层通信链路
- **部署模式升级**：支持大 EP 专家并行 + PD 分离部署，进一步提升系统吞吐性能

关键事件

2025.09

- **0day支持DeepSeek-V3.2-Exp**

2025.10

- Qwen3-MoE支持
- DeepSeek-V3.2-Exp支持W8A8C8量化
- HunyuanVideo支持Ulysses SP/ RingAttention SP / TeaCache加速
- Wan2.2-I2V支持Ulysses SP、CFG并行、VAE并行
- DeepSeek-R1/Kimi-K2支持
- DeepSeek-R1、Qwen2.5模型训练样例上线

2025.11

- **0day支持Kimi-K2-ThinkingAtlas A3,支持256K序列推理部署，原生W4A16量化**

2025.12（now）

- Qwen3-Next推理支持
  Longcat推理支持
  GPT-OSS推理支持
  Qwen-235B/32B长序列RL训练样例
- VGGT推理支持
  Hunyuan3D推理支持
  Pi0推理支持

CANN开源的下一步计划

2025.09

- **ops-transformer**
- **ops-nn**
- **ops-math**
- **ops-cv**
- **opbase**
- community
- catlass
- cann-recipes-infer
- ascend-transformer-boost
- ascend-boost-comm

2025.10

- **hixl**
- **graph-autofusion**
- sip(AscendSiPBoost项目)
- cann-recipes-train
- asnumpy(第三方贡献项目)

2025.11

- AscendC相关仓开源
- HCCL算子开源

2025.12（now）

- 图引擎开源
- HCCL框架、平台开源
- Runtime开源
- 工具链开源
- Host驱动开源

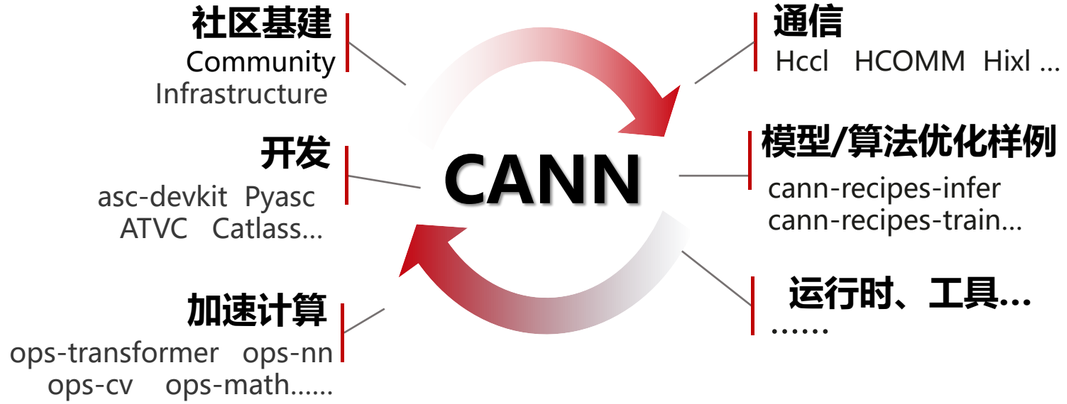
加入CANN社区

https://gitcode.com/cann

Start

STEP1

1 关注CANN社区，star感兴趣的项目



STEP2

2 了解行为准则

了解CANN社区行为准则。

STEP3

3 签署CLA

根据参与身份，选择签署CLA

个人CLA：以个人身份参与贡献，请签署个人CLA；

企业管理员：以企业管理员的身份参与贡献，请签署企业管理员CLA。

STEP4

4 参与社区共建

- **基础贡献**

参与社区会议　参与邮件讨论　提交issue/处理issue　提交PR

- **进阶贡献**

成为TSC/PMC/SIGs成员　新建SIG　组织会议　新建仓库

STEP5

5 一起成长

CANN Be X 计划	
体验官	代码&文档的bug hunting
赏金猎人	揭榜社区任务
城市主理人	成立并运营CANN城市开发者俱乐部
校园合伙人	成立并运营CANN校园开发者俱乐部
布道师	项目推广，技术与经验分享
社区建筑师	代码贡献

WIN

Thanks！

访问CANN开源社区

关注昇腾CANN公众号