CANN Meetup 北京站

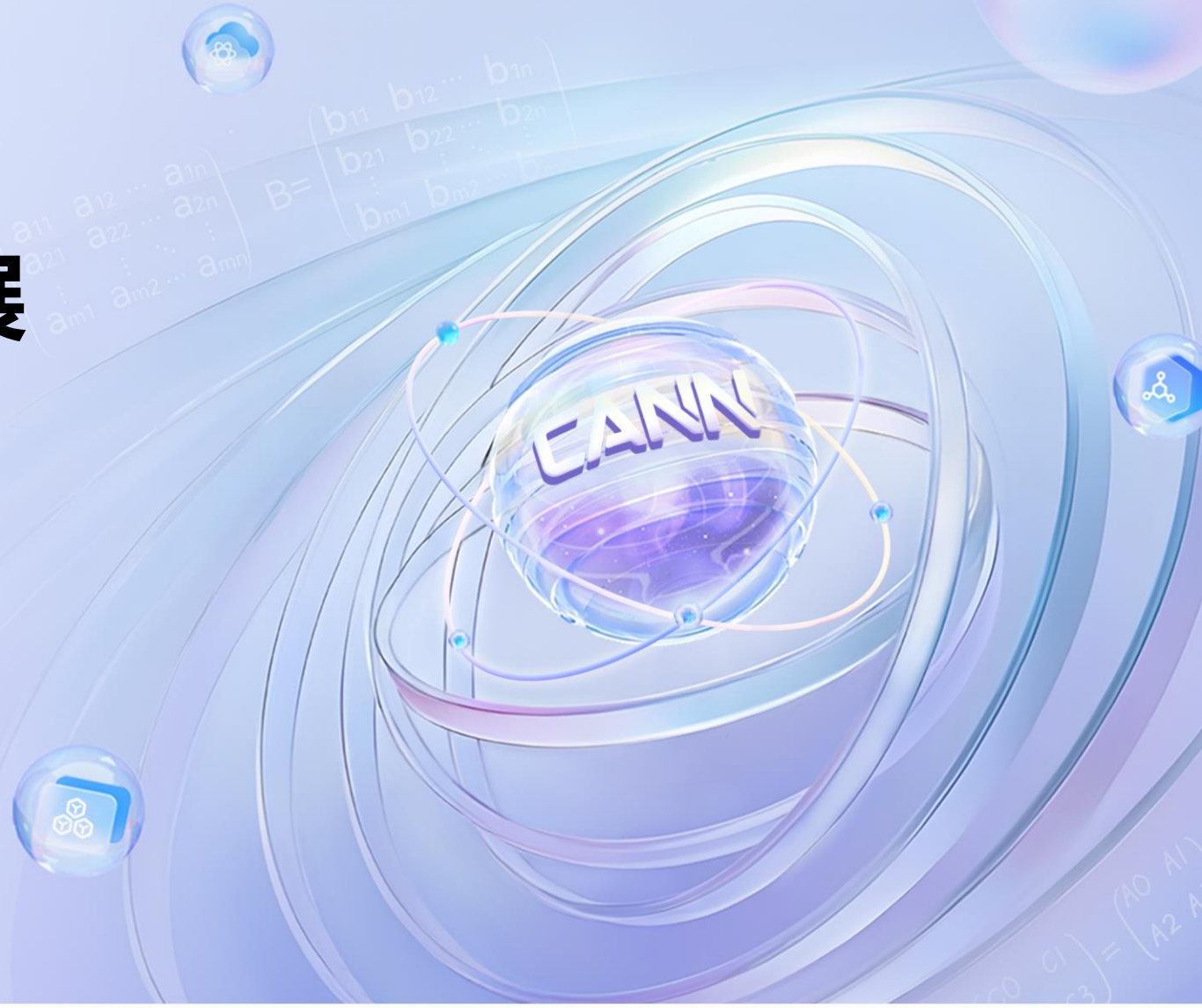
扫码关注Gitcode CANN社区

扫码关注CANN公众号

CANN社区开源进展

作者：陈敏

时间：2025年11月15日



CANN

回顾：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MLIR支持，支持Triton
- 1230 910B/910C 全面开源开放

2026

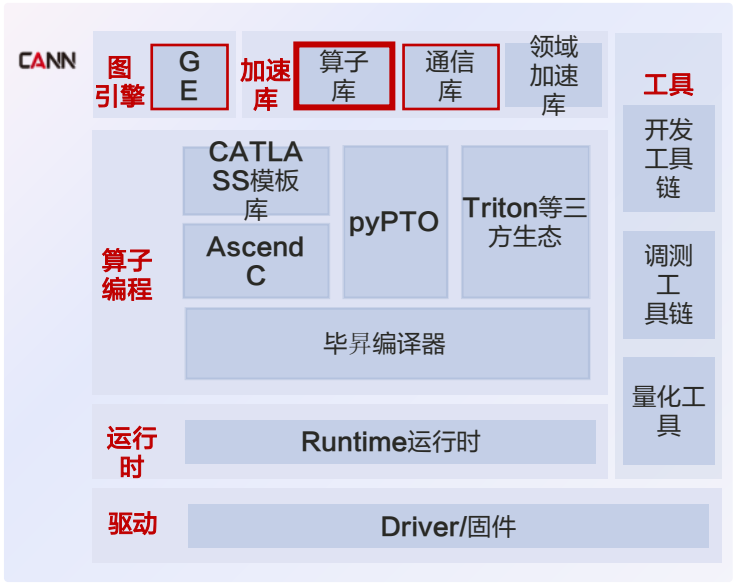
- 950系列上市即开源
- Ascend C使能下一代处理器950编程特性
- 支持多代际昇腾产品开发和创新

2027

持续迭代期：

未来每代际产品配套软件持续迭代

CANN开源进展：已开源全量算子库、部分集合通信和图引擎



https://gitcode.com/cann

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
	asc-python	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

CANN

开源试运营完成首个社区外部贡献 & 客户联创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/

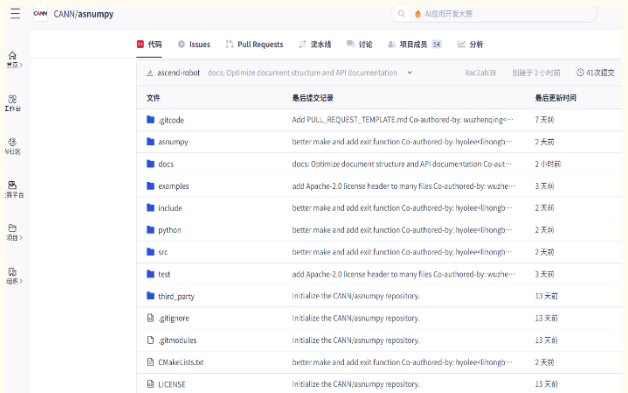
CANN 社区

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

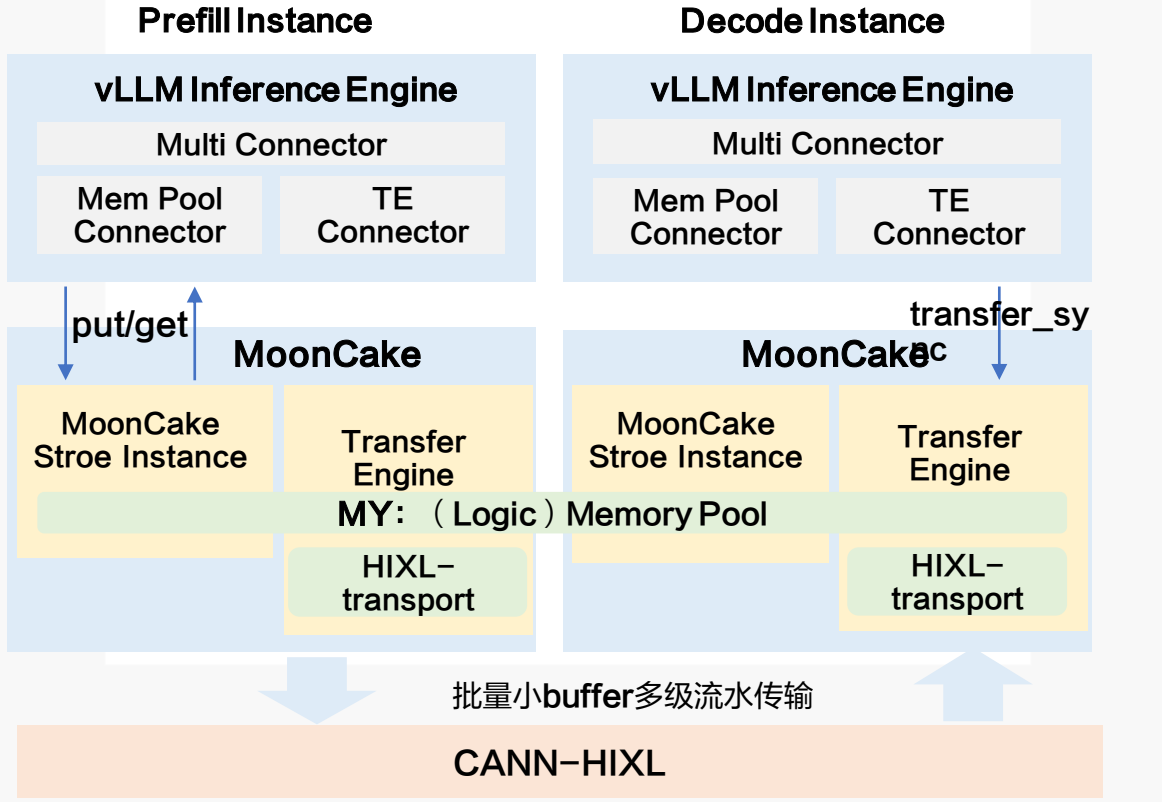


CANN-HIXL: 华为联合外部用户联创

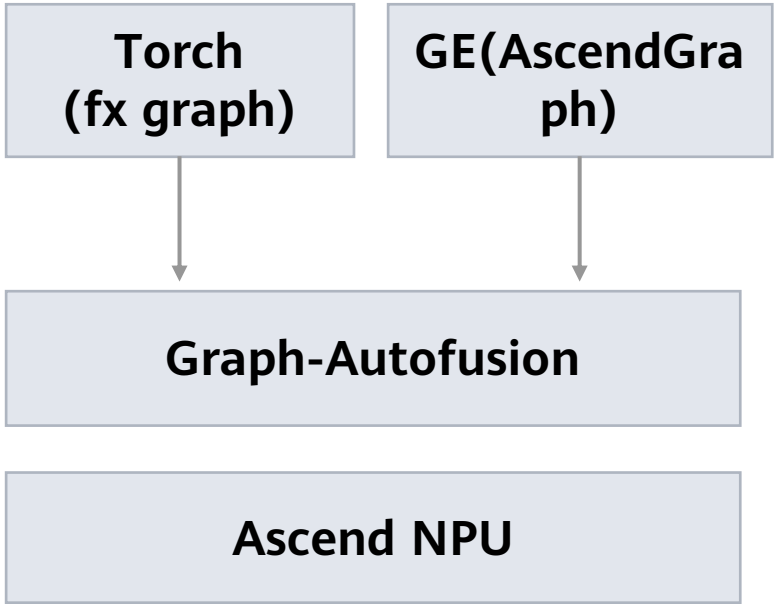
三方联合：MY + MoonCake + CANN-HIXL协同完成MY TTFT优化40%，并反哺相关优化至社区

- CANN & MoonCake：与Mooncake社区基于HIXL完成NPU对接
- MY & CANN：MY协同HIXL完成批量小Buffer多级流水传输方案在CANN的落地
- MY & MoonCake：MY贡献昇腾亲和的BatchPut/BatchGet接口至MoonCake社区

Mooncake 新动态 | 与 CANN 联合共创 HIXL 组件：开放昇腾底层高速互联，提供简易 API



图引擎自动融合仓GE-autofusion开源，支持SuperKernel



Graph-Autofusion 仓背景

- 面向昇腾芯片，提供自动融合技术
- 从GE中，将SuperKernel、自动融合codegen等融合能力抽取到Graph-Autofusion中

技术特点

- **生态对接**优先：支持 fx graph
- **轻量级**：单特性代码量数K到数十K
- **依赖少**：仅依赖 ascendc编程框架、runtime 等基础库
- **泛化性好**：基于JIT技术，在线codegen+编译

2025.10.31
- 带 SuperKernel 特性开源

2025.11.30
- 启动 inductor backend
- 启动 SuperKernel 与 ascendc 解耦工作，目标独立升级

2025.12.31
- SuperKernel 与 ascendc 可独立升级，SuperKernel 演进不再依赖升级 ascendc

2026
- SuperKernel、inductor 后端持续演进

https://gitcode.com/cann/graph-autofusion

CANN

0day 适配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适配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

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

0day 适配DeepSeek-V3.2-Exp

关键事件

2025.09

2025.10

- Qwen3-MoE适配
- 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适配
- DeepSeek-R1、Qwen2.5模型训练样例上线

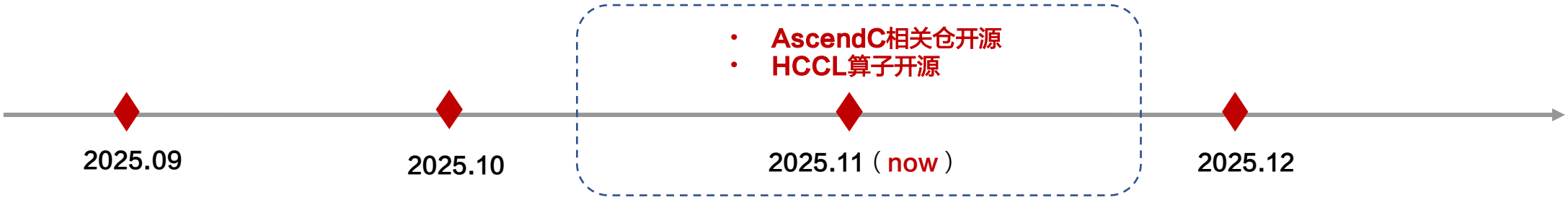
2025.11（now）

2025.12

- 敬请期待…

CANN

CANN开源的下一步计划



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

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

- 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