



天数智芯
Iluvatar CoreX

天数智芯

版本发布说明

版本：V4.0.1-MR

日期：2024.5.31

适用产品：智铠 50 | 智铠 100

1 声明

1.1 版权声明

版权所有。未经天数智芯书面许可，不得以任何形式或方式将本文档的任何部分复制，传播，转录或翻译成任何语言。

1.2 免责声明

天数智芯可以随时对本文档或本文档中描述的产品进行改进和/或更改。本文档包括与天数智芯产品有关的信息，作为说明典型应用的一种方式，因此，不一定提供足以进行生产设计的完整信息。对于本文档中内容的准确性或完整性，天数智芯不做任何陈述或保证。

1.3 联系方式

地址：上海闵行区陈行公路 2168 号 3 幢

电话：021-68886607

网址：www.iluvatar.com

Contents

1 声明	2
1.1 版权声明	2
1.2 免责声明	2
1.3 联系方式	2
2 版本发布说明	4
2.1 修订记录	4
2.2 发布 whl 包说明	4
2.3 Version 4.0.1	5
2.4 Version 4.0.0	6
2.5 Version 3.2.0	7
2.6 Version 3.1.1	8
2.7 Version Beta 2	9
2.8 Version Beta 1	10
3 商标声明	11

2 版本发布说明

2.1 修订记录

- COREX01-MR401-RN01-00: 2024/5/31

文档本次发布内容与 V4.0.0-MR 文档相比 (COREX01-MR400-RN01-00), 有以下更新:

- 新增 **Version 4.0.1**, 提供 V4.0.1-MR 的版本发布说明

2.2 发布 whl 包说明

Table 1: 发布 whl 包说明

发布 whl 包	说明	依赖 whl 包
torch	开源深度学习框架 PyTorch	无
tensorflow	端到端的开源机器学习平台	无
paddlepaddle	深度学习平台	无
onnxruntime_gpu	跨平台的推理和训练机器学习加速器	torch、tensorflow 和 igie
apex	PyTorch 扩展深度学习的加速库, 可以简化 PyTorch 中的混合精度和分布式训练	torch
auto_gptq	一个基于 GPTQ 的算法, 简单易用且拥有用户友好型接口的大语言模型量化工具包	torch
autoawq	4 位量化模型的软件包, 用于加速大模型推理	torch
bitsandbytes	CUDA 自定义函数的轻量级包装器, 特别是 8 位优化器、矩阵乘法和量化函数	torch
dali	加速库, 用于数据加载和预处理	torch 和 tensorflow
dropout_layer_norm	CUDA 扩展, 实现融合的 dropout+ 残差 +LayerNorm 算子	torch
flash_attn	一种具有 IO 感知且兼具快速、内存高效的新型注意力算法	torch 和 triton
fused_dense_lib	CUDA 扩展, 实现了融合的矩阵乘法 + 偏置 (前向和后向), 以及融合的矩阵乘法 + 偏置 +GELU(前向和后向)	torch
fused_softmax_lib	融合 softmax 算子	torch
horovod	开源的分布式训练框架, 用于扩展多加速卡训练	torch 和 tensorflow
igie	针对天数智芯加速卡研发的高性能、高通用、全流程的神经网络推理框架	torch 和 tensorflow

发布 whl 包	说明	依赖 whl 包
ixformer	针对天智智芯加速卡研发的用于加速 Transformer 的框架	torch 和 tensorflow
ixrt	天智智芯的专用推理加速引擎，支持对业界主流训练框架的模型进行解析及优化	torch
megatron_deepspeed	大型 Transformer 语言模型分布式训练框架	torch 和 triton
megatron_lm	超大 Transformer 语言模型的分布式训练框架，高效的模型并行 (张量、序列和流水线) 和多节点预训练	torch 和 triton
pcdet	一个用于点云目标检测的开源工具包	torch 和 tensorflow
pyxistream	天智智芯提供的面向 AI 智能视频分析以及多传感器数据流处理的工具包	igie 和 ixrt
rotary_emb	旋转式位置编码，可以提升 Transformer 架构的性能	torch
torch_cluster	用于 PyTorch 的高度优化的图集群算法的小型扩展库	torch
torch_quiver	用于在加速卡上进行低延迟、高吞吐量的分布式图学习的 PyTorch 库	torch 和 torch_sparse
torch_scatter	高度优化的稀疏更新 (scatter and segment) 操作扩展库，用于 PyTorch	torch
torch_sparse	优化的 Autograd 稀疏矩阵运算扩展库，用于 PyTorch	torch
torchaudio	音频信号处理的数据处理和转换，用于 PyTorch	torch 和 triton
torchvision	特定于计算机视觉的数据集、图像转换和模型，由 PyTorch 提供支持	torch 和 triton
triton	一种编写高效定制深度学习原语的语言和编译器	torch
vllm	面向大语言模型的高吞吐量和内存效率的推理和服务引擎	torch
xentropy_cuda_lib	优化的交叉熵损失，支持 bfloat16，并支持就地反向传播以节省内存	torch

2.3 Version 4.0.1

新增特性：

- 新增对 ARM 架构的支持，支持特性与 x86 架构的 V4.0.0-MR 版本支持的特性基本一致
 - 详情参考[Version 4.0.0](#)
 - ARM 架构暂不支持 GDS 基准测试工具 ixGDS

已知问题：

- x86 架构
 - JPU 解码在 jpeg 频繁创建销毁 (2 秒钟创建销毁) 的情况下，可能会有内存泄露问题，将在后续版本修复

- 解码器 rtsp 删除添加测试，每隔 1 分钟随机添加删除流，长时间运行有概率发生内存泄露，将在后续版本修复
- ARM 架构
 - 部分模型在 ARM 上出现性能下降或波动，可能受 ARM CPU 的影响，具体原因排查中，该版本暂时只提供功能支持，性能问题将在后续版本中进一步明确目标和优化提升

2.4 Version 4.0.0

推理框架与推理引擎：

- 新增适配大模型推理框架 IxFormer v0.4.0，支持 bitsandbytes FP4 推理、AutoGPTQ INT4 推理等内容
- 新增适配推理服务框架 vLLM v0.2.6，支持 BF16、Rope Scaling、AWQ 等内容
- 新增适配推理服务框架 TGI v1.1.0
- 新增适配框架 AutoAWQ v0.1.8
- 新增发布面向 AI 智能视频分析 (IVA) 以及数据流处理的 IxStream v1.0.5，并提供运行示例
- 优化升级 IxRT 推理引擎至 v0.9.0，完善了对 TensorRT-9.0 中与 TensorRT-LLM 相关 API 的兼容支持；支持了 TensorRT-8.4 以后版本的重要 API 的功能；完善了自定义网络 API 的实现
- 优化升级 IGIE 推理引擎至 v0.9.1，兼容适配 IxRT 的 TensorRT 接口 (Python 接口)
- 优化升级适配 PyTorch 框架至 v2.1.1
- 优化升级适配 PaddlePaddle 框架至 v2.5.2，并修复相关已知功能

模型与示例：

- 新增 Transformer 模型的 IxRT 自定义网络 API 的实现
- 新增提供大模型推理示例包，支持 LLaMA2-70B、ChatGLM3-6B、Baichuan2-13B、QWEN-14B 等
- 优化主流通用深度学习推理模型性能，提供 Inference Samples 模型示例

函数库：

- 优化升级 ixCCL 通信库，优化大模型推理通信相关内容
- 优化并完善高性能加速库 ixINFER、ixDNN、ixBLAS、ixFFT、CV-CUDA 等
- 优化多媒体编解码库 ixCODEC 和 ixJPEG，并修复相关已知问题

驱动：

- 优化驱动功能，修复相关已知问题

工具：

- 新增支持 GDS (GPU Direct Storage) 基准测试工具 ixGDS
- 新增支持通讯库性能测试工具 ixCCL-tests
- 优化 ixSMI 系统管理工具，适配更多硬件版本，并优化显示
- 优化 ixSYS 系统性能分析工具，并修复相关已知问题
- 优化 ixKN 核函数性能分析工具的 HBM 和 ALU 利用率实现，并修复相关已知问题

编译器：

- 优化并完善 ixCC 编译器，修复 LLVM 16 适配中的已知问题

固件升降级工具：

- 优化固件版本升级/降级工具，增加对不同板卡 Data 校验功能

文档：

- 新增《IxFormer(vLLM) 大模型推理框架使用指南》和《IxStream 流分析工具包使用指南》
- 其他更新项，请参考各文档详情

已知问题：

- 128 路 1080p 30fps 视频推 rtsp 流解码，在不保存解码文件的场景下，部分路可能有异常报错，该问题将在后续版本修复。

2.5 Version 3.2.0

Note

以下发布项若无特殊说明，x86 架构和 ARM 架构保持一致。

推理框架与推理引擎：

- IGIE 推理框架与 IxRT 推理引擎合并。IGIE 作为推理入口，新增对 IxRT Backend 的支持，新增 Fallback 机制，新增对 IxRT C++ 接口的支持。
- IxRT 新增对标 TensorRT 的核心 API 和插件支持，支持 Dynamic Shape 推理与多线程推理，新增混合精度推理接口。

函数库：

- ixInfer 支持任意尺寸 FP16 GEMM，支持任意尺寸 INT8 GEMM TCU 的使用。
- ixInfer 算子优化更新，包括 ConvTranspose 优化、小 size GEMM 优化、device select 优化、conv2d transposed channel=1 优化与 batchsize24 相关算子优化等。
- ixInfer 大幅提升 FMHA 算子的性能，包括 INT8 和 FP16 等 Attention 性能的大幅提升。

多媒体：

- JPU 完善错误处理流程，优化编解码库的设计，完善 JPU 接口的兼容性，对于小 batch size 性能有一定提升。

驱动：

- 新增多 Stream 支持，固件版本校验等功能。
- 改进进程异常终止结束策略。

工具：

- ixSYS 新增对 GPU metrics profiling 和 NCCL 的 profiling 的支持，完善 NVTX 的支持。
- ixKN 完善并发布 ixKN-UI 工具。
- (仅 x86 架构支持) ixSMI 新增对 vGPU 的支持。

虚拟化：

- (仅 x86 架构支持) 新增 GPU 虚拟化 (svGPU) 功能。

2.6 Version 3.1.1

推理框架与推理引擎：

- IGIE 推理框架：
 - 增加 IGIE Perchannel 量化支持。
 - 增加 ixInfer 自动算法选择功能，build 接口增加 Precision 参数，支持根据不同精度选择相应的优化策略。
 - 增加 Runtime Debug 输出。
 - IGIE-EXEC 增加多输入支持，YOLO detect 算子自动识别增加常用数据集的前后处理功能。
- IxRT 推理引擎：
 - 增加动态 batch size，支持 GPU IO Python 接口。

模型推理：

- 新增 BERT-Base, ResNet50, YOLOv5s, Transformer, Conformer 模型的 FP16 推理示例脚本，优化推理性能。
- ONNXRuntime 增加 TvmExecutionProvider，支持 FP16/INT8 推理通过其对接 IGIE 推理框架。
- 支持飞桨原生推理库 Paddle Inference 的 FP16/INT8 推理对接 IGIE 推理框架。
- 更新了 ResNet50, VGG, YOLOv3, YOLOv5m, YOLOv7, YOLOX, Bert-Base, Bert-Large 模型的 INT8 推理示例脚本，提升推理性能。

多媒体：

- JPU 支持批量写寄存器，支持输入 YU12、NV12、NV21，支持输出 RGB、RGBI、BGR、BGRI，支持 YUV2RGB。
- 优化了 JPU 编解码性能。
- 优化了 VPU 解码性能。

函数库：

- ixInfer：
 - 新增并优化了 Resize, Fused Layernorm, FMHA, GEMM + Bias + Activate 等算子的实现。
 - 优化了 conv3d INT8/FP16 推理的性能，加速了反卷积、并行卷积、分组卷积等。
 - 优化了 INT8/FP16 小 batch 推理的性能，支持 RNN 等循环神经网络算子，支持 LSTM 等循环神经网络。

工具：

- ixSMI 工具支持 JPU 和 VPU 的编解码使用率，支持显示风扇速度。
- ixPROF 工具支持 application replay 模式。
- ixKN 工具支持 roofline，支持 Occupancy 计算功能。
- ixSYS 支持对 ixCCL, VPU, JPU 的 profiling。

安装：

- 安装包添加了 **--disable-dkms** 选项，用于选择是否禁用 DKMS 功能。

新增对 ARM 架构的支持，支持特性与 x86 架构支持的特性一致。

2.7 Version Beta 2

IGIE 推理引擎：

- 增加自动图优化特性，适配典型模型，自动匹配 ixInfer 函数库的融合算子。
- 模型推理工具 IGIE-EXEC 提供 ResNet 系列，VGG 系列，YOLO 系列模型测试；增加通用模型的测试脚本，用户可用以评估常见模型。
- 优化 dense sme, Layout_transform 等实现，优化 transpose 算子，增加 ONNXRuntime 量化异常捕获等。
- 支持 ixInfer 所有算子，充分利用 ixInfer 推理函数库的优化更新。

模型推理：

- 新增 YOLOv5m, YOLOv7, YOLOX-m, Bert-Base-CRF, Bert-Large 模型的 INT8 推理示例脚本，更新了 ResNet50, YOLOv3 推理脚本。
- 适配 ONNXRuntime v1.13。

多媒体：

- 图像编解码 (JPU) 新增 API，提供 C/C++ 编解码接口；支持单进程单核和多进程多核编解码。
- 视频硬件解码 VPU (Video Processing Unit) 模块提供全加速硬件视频解码能力，支持多路同时解码，可以将压缩格式的视频、视频流（可以是文件格式或实时流）通过本解码模块输出为多种 YUV 格式。

函数库：

- 提供 ixJPEG 函数库，为深度学习和超大规模多媒体应用程序中常用的图像格式提供高性能、GPU 加速的 JPEG 解码功能。
- 提供 ixDEC 函数库，为视频硬件解码 VPU 模块提供用于编程软件 API 与库所支持的 C++/Python API。
- 提供 ixInfer 推理函数库：
 - 扩充并优化算子 cropAndResizePlugin, Maxpooling, Concat 等性能。
 - 优化 Conv/GEMM/Group Conv/Depth Conv，性能普遍提升。
 - 在小 batch 情况下性能普遍提升。
- 新增 ixBlasLt API 支持。

工具：

- ixSMI 工具增加显示解码利用率功能。
- 新增 ixKN-UI 工具，增加 UI 对多个进程场景的支持能力，支持 8 个 sections 和 baseline 的功能，支持将模型中的多个 kernel 存到一个数据库文件，持续的优化和完善 UI 的显示。
- ixKN 工具优化 event profiling 的流程，支持 ComputeWorkloadAnalysis, LaunchStats, SpeedOfLight, WarpStateStats, Instruction, SchedulerStats, Memory, Occupancy 的 section 数据分析以及基础的 kernel 性能分析优化提示功能。

安装：

- 增加 corex-driver.run 包，用于独立安装驱动。
- 优化了安装过程中的终端 UI 交互。

2.8 Version Beta 1

- 支持主流编程框架：PyTorch 1.10 和 TensorFlow 2.5 推理。
- 提供针对智铠加速卡适配的 IGIE 推理引擎。
- 提供视频解码支持的 C++/Python API。
- 新增了 VPU FFmpeg 视频解码接口方案。
- 提供针对智铠加速卡适配的 ixInfer 推理函数库，支持基于 INT8 数据类型在包括 IGIE 和其它框架中用于推理的 API。
- 提供针对智铠加速卡优化的 ixBLAS 和 ixDNN 函数库。
- 提供针对智铠加速卡适配的编译器、驱动和运行时。
- 提供系统监控查询和管理工具 ixSMI。
- 提供针对智铠加速卡的硬件测试工具，包括带宽测试工具 BandwidthTest、矩阵算力测试工具 ixGEMM (INT8, FP16, FP32)、P2P 性能测试工具 P2pBandwidthTest 等。
- 支持主流通用深度学习推理模型，支持 FP32/FP16/INT8 推理，并提供 VGG16、ResNet50、BERT-Base、Tacotron2、YOLOv3 网络模型示例脚本。
- 兼容 CUDA。
- 支持 Ubuntu 18.04, CentOS 7 等操作系统。

3 商标声明

- 天数智芯、天数智芯 logo、Iluvatar CoreX 等商标、标识、组合商标为上海天数智芯半导体有限公司之注册商标或商标，受法律保护。
- 除了天数智芯的注册商标外，本内容中使用的其他产品名称及标志仅用于识别目的，该等名称及标志可能是归属于其各自公司的商标。我们否认对该等名称及标志的所有权利。
- CentOS 标识为 Red Hat 公司的商标。
- Docker 为 Docker 公司在美国和其他国家的商标或注册商标。
- Linux 为 Linus Torvalds 在美国和其它国家的注册商标。
- NVIDIA 和 CUDA 为 NVIDIA 公司在美国和/或其它国家的商标和/或注册商标。
- PyTorch 为 Facebook 公司的商标。
- TensorFlow 为 Google 公司的商标。
- Ubuntu 为 Canonical 公司的注册商标。