AI 应用发布说明

1. 概述

本文档包含了此次发布的 AI 应用软件包的特性，已知问题和使用限制等。

此次发布是 MXC500 GPU 的 AI 应用交付，**MXMACA-C500-SDK-2.25.0.7**，**MXMACA-C500-Driver-2.25.0.3**，**MXMACA-C500-Pytorch-2.25.0.0** 版本适用于曦云®C500、C280、C290 OAM1.5、C290 OAM2.0、C550 OAM1.5、C550 OAM2.0 和曦思®N260。

表 1 列出了系统测试覆盖率和通过率。

表 1 系统测试覆盖率及通过率

系统	测试覆盖率/通过率
mxc500-ppl.llm.serving-2.25.0.5	Full regression with release quality
mxc500-modelzoo.llm.ppl-2.25.0.5	Full regression with release quality
mxc500-vllm-2.25.0.6	Full regression with release quality
mxc500-modelzoo.llm.vllm-2.25.0.6	Full regression with release quality

1.1 交付内容

此次发布包含以下内容：

- 二进制文件，动态库和容器镜像文件
- PPL LLM 大模型推理引擎
- vLLM 大模型推理框架
- 《曦云®系列通用计算 GPU AI 推理用户手册》

2. 新增特性及变更

本章列出历次发布的新增特性及变更。

2.1 配套 MXMACA-C500-SDK-2.25.0.7，MXMACA-C500-Driver-2.25.0.3，MXMACA-C500-Pytorch-2.25.0.0

2.1.1 mxc500-ppl.llm.serving-2.25.0.5

模块	特性说明
ppl.llm.serving	修复了部分 kernel 地址越界问题

2.1.2 mxc500-modelzoo.llm.ppl-2.25.0.5

模块	特性说明
modelzoo.llm.ppl	修复了部分 kernel 地址越界问题

2.1.3 mxc500-vllm-2.25.0.6

模块	特性说明
vllm	优化了性能
	解决部分多卡运行问题

2.1.4 mxc500-modelzoo.llm.vllm-2.25.0.6

模块	特性说明
modelzoo.llm.vllm	优化了性能
	解决部分多卡运行问题

2.2 配套 MXMACA-C500-SDK-2.24.0.12，MXMACA-C500-Driver-2.24.0.10，MXMACA-C500-Pytorch-2.24.0.5

2.2.1 mxc500-megatron-lm-2.24.0.4

模块	特性说明
Megatron-LM	优化框架性能，新增 chatglm3，qwen2，qwen1.5，baichuan2 模型支持

2.2.2 mxc500-modelzoo.llm.ppl-2.24.0.4

模块	特性说明
modelzoo.llm.ppl	新增支持模型 Baichun2-13B

2.2.3 mxc500-ppl.llm.serving-2.24.0.4

模块	特性说明
PPL-LLM	新增支持模型 Baichun2-13B

2.2.4 mxc500-vllm-2.24.0.4

模块	特性说明
vLLM	版本适配切换到 0.5.4
	优化 paged attn、fuse moe kernel 实现
	新增 Llama3.1 等模型适配

2.2.5 mxc500-modelzoo.llm.vllm-2.24.0.4

模块	特性说明
modelzoo.llm.vllm	新增兼容 vllm 0.5.4 脚本，新加若干个模型
	提供 torch profile 方式

2.2.6 mxc500-alpaca-lora-2.24.0.4

模块	特性说明
alpaca-lora	支持 alpaca-7b 和 alpaca-13b

2.2.7 mxc500-paddle-2.24.0.5

模块	特性说明
Paddle-maca	优化了部分算子性能

2.2.8 mxc500-internlm-2.24.0.4

模块	特性说明
internlm	增加对 InternEvo 的支持

2.2.9 mxc500-modelzoo.llm.diffusers-2.24.0.6

模块	特性说明
modelzoo.llm.diffusers	增加多卡运行代码

2.3 配套 MXMACA-C500-SDK-2.23.0.1018，MXMACA-C500-Driver-2.23.0.1014，MXMACA-C500-Pytorch-2.23.0.1011

2.3.1 mxc500-onnxruntime-2.23.0.3

模块	特性说明
Onnxruntime-maca	进一步优化了 conv、resize、transpose、bridge 等算子的性能
	解决了部分模型无法运行的问题

2.3.2 mxc500-ppl.llm.serving-2.23.0.3

模块	特性说明
PPL-LLM	优化了部分算子性能

2.3.3 mxc500-modelzoo.llm.ppl-2.23.0.3

模块	特性说明
modelzoo.llm.ppl	提供了若干个 ppl 测试样例

2.3.4 mxc500-vllm-2.23.0.3

模块	特性说明
vLLM	bloom 模型推理问题，修复 qwen_moe 和 deepseek 推理问题
	优化 paged attn
	新增 lora、multi-lora 支持和优化
	新增 gptq、awq 功能支持，性能待优化

2.3.5 mxc500-modelzoo.llm.vllm-2.23.0.1

模块	特性说明
modelzoo.llm.vllm	提供了若干个 vllm 测试样例，测试 dytpe 改成 float16

模块	特性说明
	新增 lora 和 gptq 测试脚本

2.3.6 mxc500-internlm-2.23.0.1

模块	特性说明
internlm	添加了训练脚本和 README 文档
	删除了不再起作用的环境变量

2.3.7 mxc500-modelzoo.llm.diffusers-2.23.0.1

模块	特性说明
modelzoo.llm.diffusers	提供了 diffusers 的 onnx 后端测试样例
	新增 Prati 数据集测试方式
	完善打印信息

2.3.8 mxc500-paddle-2.23.0.1

模块	特性说明
Paddle-maca	更新到 2.6.0 版本

2.3.9 mxc500-modelzoo.cnn.training-2.23.0.1

模块	特性说明
modelzoo.cnn.training	解决部分模型运行报错问题

2.4 配套 MXMACA-C500-2.22.0.9 amd64 和 MXMACA-C500-2.22.0.11 arm64

2.4.1 mxc500-onnxruntime-2.22.0.9.318/mxc500-onnxruntime-ft2000-2.22.0.11.159

模块	特性说明
Onnxruntime-maca	优化了部分算子的性能

2.4.2 mxc500-megatron-lm-2.22.0.9.306

模块	特性说明
Megatron-LM	优化框架性能，支持 core0.6.0

2.4.3 mxc500-ppl.llm.serving-2.22.0.9.311/mxc500-ppl.llm.serving-ft2000-2.22.0.11.168

模块	特性说明
PPl-LLM	优化了部分算子性能

2.4.4 mxc500-modelzoo.llm.ppl-2.22.0.9.118

模块	特性说明
modelzoo.llm.ppl	提供了若干个 ppl 测试样例

2.4.5 mxc500-vllm-2.22.0.9.186

模块	特性说明
vLLM	优化 gemm 计算
	修复长文本 oom 问题

2.4.6 mxc500-modelzoo.llm.vllm-2.22.0.9.122

模块	特性说明
modelzoo.llm.vllm	提供了若干个 vLLM 测试样例

2.4.7 mxc500-internlm-2.22.0.9.33

模块	特性说明
internlm	internlm 大模型训练框架

2.4.8 mxc500-modelzoo.llm.diffusers-2.22.0.9.120

模块	特性说明
modelzoo.llm.diffusers	提供了 diffusers 的 onnx 后端测试样例

2.4.9 mxc500-modelzoo.llm.transformers-2.22.0.9.115

模块	特性说明
transformers	提供了 transformers 测试环境和测试代码

2.4.10 mxc500-modelzoo.cnn.training-2.22.0.9.61

模块	特性说明
modelzoo.cnn.training	提供了若干个 cnn training 测试样例

2.4.11 mxc500-modelzoo.cnn.inference-2.22.0.9.120

模块	特性说明
modelzoo.cnn.inference	提供了常见的 CNN 模型的 ONNXRUNTIME 推理测试样例

2.4.12 mxc500-bitsandbytes-2.22.0.9.150

模块	特性说明
bitsandbytes	支持了部分场景下 bitsandbytes 的 API

2.5 配套 MXMACA-C500-2.20.2.19

2.5.1 mxc500-onnxruntime-2.20.2.18.238

模块	特性说明
Onnxruntime-maca	优化了部分算子性能
	解决了部分模型无法运行的问题

2.5.2 mxc500-megatron-lm-2.20.2.2.141

模块	特性说明
Megatron-LM	新增对 megatron-Core 的支持

2.5.3 mxc500-ppl.llm.serving-2.20.2.18.236

模块	特性说明
PPl-LLM	新增支持了 Qwen、Mixtral、Llama3 模型
	优化了部分模型性能

2.5.4 mxc500-colossalai-2.20.2.2.91

模块	特性说明
ColossalAI	首次发布，优化 optimizer 性能

2.5.5 mxc500-vllm-2.20.2.19.147

模块	特性说明
vLLM	首次发布，兼容官方 0.4.0 版本
	存在以下局限性： ● 当前暂不支持 Lora，后续将完善支持 ● 支持 GPTQ 量化方式，暂不支持其他量化方式 ● 暂不支持 `enforce_eager=False` 方式，内部关闭 ● 暂不支持 FP8 类型的 KV Cache ● 当前仅包含 Ubuntu 20 系统版本，后续将完善支持其他系统

2.6 配套 MXMACA-C500-2.19.2.23

2.6.1 mxc500-onnxruntime-2.19.2.23.111

模块	特性说明
Onnxruntime-maca	优化部分模型计算性能

2.6.2 mxc500-ppl.llm.serving-2.19.2.23.111

模块	特性说明
PPL-LLM	优化部分模型计算性能

2.7 配套 MXMACA-C500-2.19.2.7

2.7.1 mxc500-onnxruntime-2.19.2.5.65

模块	特性说明
Onnxruntime-maca	新增 profiling 功能
	优化部分模型计算性能

2.7.2 mxc500-ppl.llm.serving-2.19.2.7.66

模块	特性说明
PPL-LLM	优化部分模型计算性能

2.8 配套 MXMACA-C500-2.19.0.12

2.8.1 mxc500-onnxruntime-2.19.0.12.40

模块	特性说明
Onnxruntime-maca	优化部分模型计算性能

2.8.2 mxc500-ppl.llm.serving-2.19.0.12.43

模块	特性说明
PPL LLM	新增支持 Llamav2，ChatGLM2，ChatGLM3 模型
	优化模型转换及服务化部署示例

2.9 配套 MXMACA-C500-2.18.0.4

无新增和变更特性，修复 reported bug。

2.10 配套 MXMACA-C500-2.17.3.11

2.10.1 mxc500-onnxruntime-2.17.3-0

模块	特性说明
Onnxruntime-maca	优化部分模型计算性能
	修复部分算子计算逻辑问题

2.10.2 mxc500-deepspeed-2.17.3.11.76

模块	特性说明
DeepSpeed	Alpha 版本，支持大模型训练

2.10.3 mxc500-megatron-lm-2.17.3.11.35

模块	特性说明

模块	特性说明
Megatron-LM	Alpha 版本，支持大模型训练

2.10.4 mxc500-paddle-2.17.3.9.111

模块	特性说明
Paddle-maca	Alpha 版本，支持 FP32 精度下的单卡及多卡训练

2.10.5 mxc500-ppl.llm.serving-2.17.3.11.58

模块	特性说明
PPL LLM	Alpha 版本，支持 LLama v1，Baichuan，InternLM 模型

2.11 MXC500-ONNXRUNTIME-2.16.1-3

模块	特性说明
Onnxruntime-maca	新增支持部分模型
	优化部分算子性能

2.12 MXC500-ONNXRUNTIME-2.15.0-4

模块	特性说明
Onnxruntime-maca	支持 C、C++和 Python 接口
	支持多种模型数据类型，包括 float32、float16、int8、uint8 等
	支持动态 batch 推理功能
	支持多线程调用和多进程调用
	支持单机多 GPU 卡
	支持用户管理系统内存、锁页内存、显存
MacaConverter	支持 Caffe、Tensorflow、Pytorch、PaddlePaddle、Darknet 模型转为 ONNX 模型
	支持 ONNX 简化
	支持 FP32 模型转为 FP16 模型
	支持子图提取、图优化
MacaQuantizer	支持多种量化算法
	支持开启强制优化
	支持量化损失阈值配置

模块	特性说明
	支持自定义预处理
	支持自动量化流程
	支持 Debug 模式
MacaPrecision	支持 MXC500 与 CPU 精度对比
	支持逐层精度对比
	支持多种精度评估方法
	支持量化模型精度分析

3. 已知问题和使用限制

模块	问题和限制说明
Paddle-maca	个别模型偶现训练报错
Onnxruntime-maca	个别模型推理中出现精度不符合预期，推理报错问题
vLLM	个别模型性能测试不稳定
	OpenAI 测试如遇问题请参考官方 issue : https://github.com/vllm-project/vllm/issues/7246
	多卡如遇 dmesg 显存超出信息为正常输出
PPL LLM	个别模型推理偶现异常
ColossalAI	如果出现 OOM： 1. 在物理机上执行 `sudo modprobe -r metax && sudo modprobe metax xcore_page_size=9` 2. 在运行命令前执行 `export MALLOC_THRESHOLD=99`
BitsAndBytes	Int8xInt8 to Int8 个别大矩阵乘法运算会有异常
	Int8xInt8 to Int8 性能较低
	Matmult 矩阵乘性能可能不稳定
modelzoo.cnn.training	ARM 平台暂不支持通过内置 dockerfile 文件来构建 modelzoo 镜像
	Pytorch 训练 centernet_R18 和 Retinanet 模型时，存在 amp 精度 loss 为 NaN 的情况
	Pytorch 训练多 VF 场景下偶发 hang
	Pytorch 训练学习率策略，推荐使用--auto-scale-lr 自适应学习率
	GPU 占用率低时受到其他硬件因素影响较大，在不同机器测试时易出现性能波动
	Pytorch 个别模型对 CPU 资源敏感易出现性能波动现象
	Pytorch ssd 模型多卡训练偶发 loss 为 NaN
	Pytorch Deeplabv3 模型 FP32 精度单卡训练时，需要设置新的环境变量以避免 loss 为 NaN
TensorFlow2	个别模型推理精度不符合预期

声明