MXMACA-C500 发布说明

1. 概述

本文档包含了此次发布的 MXMACA 驱动软件包的特性，已知问题和使用限制等。

此次发布是 MXC500 GPU 的 **MXMACA-C500-SDK-2.25.0.7**，**MXMACA-C500-Driver-2.25.0.3**，**MXMACA-C500-Pytorch-2.25.0.0**，**VBIOS 1.16.3.0** 版本交付，适用于曦云®C500、C500X、C280、C290 OAM1.5、C290 OAM2.0、C550 OAM1.5、C550 OAM2.0、ARM 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
MXMACA-C500-SDK-2.25.0.7	Full regression with release quality
MXMACA-C500-Driver-2.25.0.3	Full regression with release quality
MXMACA-C500-Pytorch-2.25.0.0	Full regression with release quality

1.1 交付内容

此次发布的软件包包含以下内容：

- MXMACA-C500-SDK-2.25.0.7 SDK 软件栈
- MXMACA-C500-Driver-2.25.0.3 Driver 软件栈
- MXMACA-C500-Pytorch-2.25.0.0 Pytorch 软件栈

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

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

模块		特性说明
MXMACA		本版本旨在快速提供 MXMACA 软件栈在 MOE（混合专家模型）技术上的最新支持成果，后续版本仍会持续优化
	mcpti	新增 Graph API IB 模式的 tracer 功能支持
	mcBLAS	增加了 Group GEMM 相关 API 的支持
		针对大模型推理相关模型进行了性能提升
		加强使用 Graph API 测试场景覆盖
	Compiler	使能 OpenAcc P0 特性和 OpenCL 基本特性
		优化 pkfma 和 FP16 cvt 执行策略，提升 vLLM 关键 Kernel GPTQ 性能
		优化冗余的跨基本块的 FP16 数据合并操作，提升 triton MMA 场景性能
		加强 Direct Dispatch 的测试场景覆盖
		加强显存使用复杂场景的测试场景覆盖
		FlashAttention 优化了 headdim96 的前向和反向功能、headdim256 的 decode 性能
		提升 FP16 MMA on triton 性能

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

模块		特性说明
UMD		提供 DirectDispatch 功能
		加强了多进程支持的稳定性，并有小幅性能提升
MCCL		支持易构集群
	C500X	Ring 算法支持网卡和 PCIe 并行通信，分布算法性能优化，TP8 带宽性能提升
	ARM	支持复用 PCIe 链路通信，单机多卡通信带宽性能提升
Graph	API	支持 Memory Node 基本功能

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

模块		特性说明
ACL	mcTracer	支持根据 UMD memory tracing log 单独生成 trace 文件，并可以通过 mcTracer-Viewer 打开并显示
	mcDNN	增加了 FP16 前向 depthwise 卷积融合功能
	mcDNN/ mcBLAS	增加外置 kernel 选择优化工具
	Flash Attention	增加 MHA/GQA backward 全部 headdim 的支持
		增加对 decoder attention 和 paged attention 全部 headdim 的支持
		支持更通用的 attention mask
Compiler		增加 global load/store builtin function with predicator
ARM		修复了一些 ARM 平台上的软件适配问题

2.3 MXMACA-C500-SDK-2.23.0.1018，MXMACA-C500-Driver-2.23.0.1014，MXMACA-C500-Pytorch-2.23.0.1011，MXMACA-C500-K8s-0.7.13

模块		特性说明
C500X	MetaXLink	支持隐式 MetaXLink training
	MCCL	支持 C500X
ACL	mcDNN	提升 FP16 depth-wise 卷积性能
	mcBLAS	提升大语言模型场景下的矩阵乘法性能
	Flash Attention	提升 head dimension 部分性能
Pytorch		新增支持 python 3.10
		新增支持 torch2.1
Compiler		新增编译选项`-mllvm -metaxgpu-lduB16=true`
Triton		支持 triton2.1
mcTracer		支持根据热点 API slice 排序
Bug 修复		修复了 200+ reported bug，包括 5+ hot issue

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

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

模块		特性说明
OS 适配		本次发布新增 OS BCLinux R8 U2，kernel 4.19.0-240.23.11.el8_2.bclinux.x86_64
		支持飞腾 5000C ARM 系统，kernel 5.15.0-1.10.6.v2307.ky10h.aarch64
驱动	Warm Reset	支持 Warm Reset 方式
ACL		发布 mcApex 和 mcXformer
Bug 修复		修复了 100+ reported bug，包括 1 hot issue

2.5 MXMACA-C500-2.20.2.19

模块		特性说明
OS 适配		本次发布新增 OS ALinux3，kernel 5.10.134-13.1.al8.x86_64
		本次发布新增 OS CTYun 23.01，kernel 5.10.0-136.12.0.86.ctl3.x86_64
		本次发布新增 OS x86_64 Kylin V10 SP2，kernel 5.10.0-136.12.0.86.ctl3.x86_64
		本次发布新增 OS KeyarchOS 5.8，kernel 4.19.91-27.4.19.kos5.x86_64
驱动	VPU	新增支持多进程 FFmpeg 编解码
Bug 修复		修复了 200+ reported bug，包括 1 hot issue

2.6 MXMACA-C500-2.19.2.23

模块		特性说明
Bug 修复		修复了 220+ reported bug，包括 19 hot issue

2.7 MXMACA-C500-2.19.2.7

模块		特性说明
Bug 修复		修复了 160+ reported bug，包括 13 hot issue

2.8 MXMACA-C500-2.19.0.12

模块		特性说明
Bug 修复		修复了 160+ reported bug，包括 7 hot issue

沐曦信息 MetaX Confidential 2024-11-18 15:00:00

2.9 MXMACA-C500-2.18.0.4

模块		特性说明
Bug 修复		修复了 reported bug

2.10 MXMACA-C500-2.17.3.11

模块		特性说明
驱动	内核态驱动	新增支持 CC Linux 22.09
	虚拟化	新增支持虚拟化相关功能
	VPU	新增支持 264/265/jpeg 编码，264/265/av1/avs2/jpeg 解码，8k 30fps
Bug 修复		修复了 57 reported bug，包括 24 hot issue

2.11 MXMACA-C500-2.16.1.11

模块		特性说明
驱动	内核态驱动	新增支持 RedHat 9/CentOS 9 和 BC-Linux for Euler（21.10 及以上）
Bug 修复		修复了 120+ reported bug，包括 10+ hot issue

2.12 MXMACA-C500-2.15.0.7

模块		特性说明
驱动	固件	支持 C500 芯片上运行的基本固件功能
	内核态驱动	支持 C500 芯片上运行的基本内核功能
	用户态驱动	支持 C500 芯片上运行的基本用户态驱动功能
编译器	编译器	支持基本 C500 编译器功能
数学库	基本数学库	支持 C500 芯片上运行的基本数学库功能
	Pytorch	支持 C500 芯片上运行的 Pytorch 功能

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

3. 已知问题和使用限制

模块	问题和限制说明
MXMACA 软件栈	多程序长时间并行执行，有概率会出现程序执行无响应问题
	Flash Attention C++接口有优化，使用 Flash Attention C++接口的应用需要继承新的接口，对 2.23.0.x 及以前版本不兼容
	某些主板不支持 atomic 访问的部分功能
	mx C/C++程序不支持在 Ubuntu18.04 系统编译运行
	特定 case 在 ARM 会遇到长时间执行无法结束的问题
模型推理和训练	在开启 cuda graph 的时候抓取 torch profile 会造成系统异常，需要配置如下环境变量规避该问题：MCPTI_ENABLED=ON 和 MACA_GRAPH_LAUNCH_MODE=1
	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
	对于 LLM.PPL，不建议开启 ACS，因为开启 ACS 可能会导致性能下降。如果必须开启 ACS，需要关闭通信库 PCIe 复用功能，环境变量是 export MCCL_PCIE_BUFFER_MODE=0
	通讯库性能优化使用了更多显存，特殊情况下，可能导致显存不足
VPU	VPU 最大只能申请到 16G HBM
	SDK 解码路数过多会有异常
	SDK 解码 AVS2 可能会有异常
	SDK 解码 JPEG 性能不达标
	编码性能可能不达标
	MCJPEG、JPEG 编解码部分分辨率会有异常 解码暂不支持 444、422、gray、411、440 格式 jpeg 文件
	FFmpeg 黑白视频解码可能有花屏
	开启虚拟化多于 1VF 情形下，编码相关功能不可用
	ARM 麒麟系统 SDK 编码和解码 jpeg 性能不达标
mx-smi	个别机型执行 warm reset 会引发异常

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

模块	问题和限制说明
	warmreset 期间不支持查看卡
	暂不支持 mx-smi --show-clocks-throttle-reason 功能
MetaXLink	MetaXLink 卡间互联后不支持透传单卡到虚拟机中使用，否则，可能会发生不可预期的行为
虚拟化	虚拟机透传 GPU 或 VF，只支持虚拟机启动前设置透传 GPU 或 VF
	/etc/mvgvm_config 文件默认权限为 0，如需要修改此文件，需要将文件权限改为 0644
	不支持 mx-smi --vfflr 操作

飞腾信息 MetaX Confidential 2024-11-18 15:00:00

声明

版权所有 ©2023-2024 沐曦集成电路（上海）有限公司。保留所有权利。

本文档中呈现的信息属于沐曦集成电路（上海）有限公司和/或其附属公司（以下统称为“沐曦”），非经沐曦事先书面许可，任何实体或个人均不得获得本文档的副本，且无权以任何方式处理本文档，包括但不限于使用、复制、修改、合并、出版、发行、销售或传播本文档的部分或全部。

本文档内容仅供参考，不提供任何形式的、明示或暗示的保证，包括但不限于对适销性、适用于任何目的和/或不侵权的保证。在任何情况下，沐曦均不对因本文档引起的、由本文档造成的、或与之相关的任何索赔、损害或其他责任负责。

沐曦保留自行决定随时更改、修改、添加或删除本文档的部分或全部的权利。沐曦保留最终解释权。

沐曦、MetaX 和其他沐曦图标是沐曦的商标。本文档中提及的所有其他商标和商品名称均为其各自所有者的财产。

飞腾信息 MetaX Confidential 2024-11-18 15:00:00