CANN
6.3.RC2
术语和缩略语

文档版本	01
发布日期	2023-07-25

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 术语和缩略语

A-E

表 1-1

术语/缩略语	含义
A	
AccumulatedRelativeError	累积相对误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大，值越接近于0，表明越相近，值越大，表明差距越大。
Advisor	专家系统。 用于聚焦模型和算子的性能调优Top问题，识别性能瓶颈，重点构建模型和算子瓶颈分析并提供优化推荐，支撑开发效率提升的工具。
AscendCL	Ascend Computing Language，昇腾编程语言。 提供Device管理、Context管理、Stream管理、内存管理、模型加载与执行、算子加载与执行、媒体数据处理等C语言的API库供用户开发深度神经网络应用，用于实现目标识别、图像分类等功能。
ADK	Ascend Development Kit，解决方案提供的开发者套件包。 通过安装相关软件包后获得开发必需的API、库、工具链等开发组件。
AI	Artificial Intelligence，人工智能。 研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。
AIPP	Artificial Intelligence Pre-Processing，AI预处理。AIPP用于在AI Core上完成图像预处理，包括改变图像尺寸、色域转换（转换图像格式）、减均值/乘系数（改变图像像素），数据处理之后再进行真正的模型推理。
AOE	Ascend Optimization Engine，昇腾调优引擎。 用于封装调优过程中的ATC编译及AscendCL运行服务接口，提供并行调优功能。

术语/缩略语	含义
Ascend EP	Ascend Endpoint昇腾AI处理器作为终端节点（从控节点）。 主要功能是配合主设备（X86，ARM等各种Server），快速高效的处理推理、训练、图像识别等工作，例如PCIe加速卡。
Ascend RC	Ascend Root Complex，昇腾AI处理器作为根组件（主控节点）。 提供主机控制功能，主要应用于移动端侧，例如Atlas 200 DK。
ATC	Ascend Tensor Compiler，昇腾张量编译器。 ● 通过ATC，可以将开源框架的网络模型（如Caffe、TensorFlow等）转换成昇腾AI处理器支持的离线模型。模型转换过程中可以实现算子调度的优化、权值数据重排、内存使用优化等 ● 通过ATC，可以进行算子编译。
AMP	Automatic Mixed Precision，自动混合精度训练。 AMP模块是PyTorch1.8.1及以上版本框架内置的使能混合精度训练的功能模块。混合精度训练是在训练时混合使用单精度（float32）与半精度(float16)数据类型，将两者结合在一起，并使用相同的超参数实现了与float32几乎相同的精度。
B	
Batch	模型训练的一次迭代（即一次梯度更新）中使用的样本集。
Batch size	模型迭代一次，使用的样本集的大小。
BIOS	Basic Input Output System，基本输入输出系统。 存于计算机主板上的一种固件。包括基本输入输出控制程序、上电自检程序、系统启动自举程序、系统设置信息，为计算机提供底层的硬件设置和控制功能
BIU	Bus Interface Unit，总线接口单元。 记录AICore和DDR/L2之间的内存访问情况
BTBC	Board-to-Board Connector，板对板连接器。 BTB连接器用于连接印刷电路板（PCB）。
C	
CAN	Controller Area Network，控制区域网络。 CAN是一种串行通讯总线，能有效地支持具有很高安全等级的分布式实时控制。
CANN	Compute Architecture for Neural Networks，AI异构计算架构。 CANN是针对AI场景推出的异构并行计算架构，通过提供多层次的编程接口，支持用户快速构建基于Ascend平台的AI应用和业务。
CCE	Cube-based Computing Engine，基于Cube的计算引擎。 CCE加速库通过API的方式，为上层应用（为机器学习提供的各种framework或者Application）提供加速。

术语/缩略语	含义
CCEC	CCE Compiler，CCE编译器。 CCE Compiler是一个异构系统编译器，是CCE异构编程语言的编译工具，它编译CCE混合代码：CCE Host代码和CCE AICPU|AICORE设备代码，产生可以在CCE系统上运行的可执行文件。
CFM	Cubic Feet Per Minute，立方英尺每分钟。 测量气体流速的单位。
CNN	Convolutional Neural Network，卷积神经网络。 是一种前馈神经网络，人工神经元可以响应周围单元，可以进行大型图像处理。
CosineSimilarity	余弦相似度算法。 精度比对算法之一，计算结果取值范围为[-1,1]，比对的结果如果越接近1，表示两者的值越相近，越接近-1意味着两者的值越相反。
CPU	Central Processing Unit，中央处理单元。 是计算机的主要设备之一，其功能是解释计算机指令以及处理计算机软件中的数据，与内部存储器、输入及输出设备成为现代电脑的三大部件。
Cube	Cube是AI Core中的运算单元，主要处理矩阵乘积累加运算。
D	
DDP	Distributed Data Parallel，分布式训练。 同时利用一台或者多台机器上的GPU/NPU进行并行计算。
DDR	Double Data Rate，双倍数据速率。 与传统的单数据速率相比，DDR技术实现了一个时钟周期内进行两次读/写操作，即在时钟的上升沿和下降沿分别执行一次读/写操作。
DL	Deep Learning，深度学习。 是机器学习的分支，是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。
DSMI	Device System Manage Interface，设备系统管理接口。
DVPP	Digital Vision Pre-Processing，数字视觉预处理。 提供对特定格式的视频和图像的进行解码、缩放等预处理操作，以及对处理后的视频、图像进行编码再输出的能力。
单算子比对	精度比对工具里Tensor比对的一种方式。选择网络模型中一个或多个参与计算的算子进行精度比对。
E	
ECC	Error Checking and Correction，错误检查和纠错技术。 该技术在原来的数据位中增加校验位，通过检验位来检测数据错误，并具备错误纠正能力。

术语/缩略语	含义
EMMC	Embedded Multimedia Card，多媒体存储卡。 是一种新的存储技术，外部提供的接口和SD卡接口类似，内部存储介质为Flash，并且内建坏块管理系统。
EP	Endpoint，终端节点。 EP是具有PCIe接口的网卡、SATA控制器等。
Epoch	数据集的一次完整遍历。
EVB	Evaluation board，评估板。 EVB板用于芯片的性能、可靠性、集成测试的验证。

F-J

表 1-2

术语/缩略语	含义
F	
FE	Fusion Engine，融合引擎。 提供图优化，图编译实现接口； 实现算子接入管理；实现算子融合优化。
FLOPS	Floating-Point Operations Per Second，每秒所执行的浮点运算次数。 常被用来估算电脑的执行效能，尤其是在使用到大量浮点运算的科学计算领域中。正因为FLOPS字尾的那个S，代表秒，而不是复数，所以不能省略掉。
Framework	NN框架执行引擎。 包括开源主流框架TensorFlow、PyTorch、Caffe等，自研框架MindSpore。系统针对开源主流框架提供适配插件，从而可利用昇腾AI处理器加速计算能力。
G	
GDAT	Gradient Auto Tuning，梯度调优。 GDAT是通过最大化反向计算与梯度聚合通信并行度，缩短通信拖尾时间的优化工具。分布式训练场景下，各个设备之间计算梯度后执行梯度聚合操作，梯度聚合算子的融合策略会影响反向计算结束后的通信拖尾时间，从而影响集群训练的性能和线性度。
GDB	GNU Debugger，GNU调试器。 UNIX及UNIX-like下的命令行调试工具，可以执行程序、管理断点、检查变量赋值、调用函数等。

术语/缩略语	含义
GE	Graph Engine，图引擎。 提供了Graph/Operator IR作为安全易用的构图接口集合，用户可以调用这些接口构建网络模型，设置模型所包含的图、图内的算子、以及模型和算子的属性。
GPU	Graphics Processing Unit，图形处理器。 GPU是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。
Graph模式	MindSpore的静态图模式，将神经网络模型编译成一整张图，然后下发执行，性能高。
H	
HCC	Huawei Compiler Collection，华为编译器。
HBM	High Bandwidth Memory，高带宽存储器。 高带宽存储器是超微半导体和SK Hynix发起的一种基于3D堆栈工艺的高性能DRAM（dynamic random access memory），适用于高内存带宽需求的应用场合，像是图形处理器、网络交换及转发设备（如路由器、交换器）等。
HCCL	Huawei Collective Communication Library，华为集合通信库。 HCCL提供了深度学习训练场景中服务器间高性能集合通信的功能。
HDC	Host Device Communication，主机设备通信。 用于Host和Device之间通信模块，在Host和Device里面均有部署。
HDR	High Dynamic Range，高动态范围。 摄影术语，用来描述媒体应用，如数字影像和数字音频。
HWTS	Hardware Task Scheduler，硬件任务调度。 提供对AI Core任务的硬件调度能力，减少调度时延。
I	
I2C	Inter-integrated Circuit，集成电路总线。 I2C总线允许在同一电路板上的组件之间轻松通信。
IDE	Integrated Development Environment，集成开发环境。
IFU	Instruction Fetch Unit，取指单元。 记录每一次icache访问情况
IPC	IP Camera，IP摄像机。
IR	Intermediate Representation，中间表示。 IR是一种数据结构，可将输入的资料建构为一个计算机程序，也可以将一部分或是所有输出的程式反推回输入资料。

术语/缩略语	含义
ISP	Image Signal Processing，图像信号处理。 主要用来对前端图像传感器输出信号处理的单元，以匹配不同厂商的图象传感器。
IVS	Intelligent Video Surveillance，智能视频采集系统。 提供集管理、存储、智能分析及应用、编解码为一体的智能视频采集系统。
IMU	I/O Board Management Unit，IO板管理单元。
J	
JPEGD	JPEG Decoder，JPEG图像解码器。 提供对JPEG格式的图像进行解码的能力。
JPEGE	JPEG Encoder，JPEG图像编码器。 提供对图像进行编码输出为JPEG格式的能力。
精度比对	通过NPU运行生成的dump数据与Ground Truth（基于GPU/CPU运行生成的npy数据）进行比对。实现自主研发算子与业界标准算子运算结果的差异比较。

K-O

表 1-3

术语/缩略语	含义
K	
KullbackLeiblerDivergence	KL散度算法。 精度比对算法之一，计算结果取值范围为0到无穷大。KL散度越小，真实分布与近似分布之间的匹配越好。
L	
LAN	Local Area Network，局域网。 由处于同一建筑或方圆几公里范围内的个人计算机和工作站相连接而组成的网络，具有高速和低错误率的特点。
LLC	Last Level Cache，最后一级Cache。 在访问内存之前调用的共享最高级别缓存通常称为最后一级缓存（LLC）。
Loss	损失，预测值与实际值的偏差，深度学习用于判断模型好坏的一个标准。

术语/缩略语	含义
LPDDR4x	Low-Power DDR4x，低功耗内存技术。 面向低功耗内存而制定的通信标准，以低功耗和小体积著称，专门用于移动式电子产品。
M	
MaxAbsoluteError	最大绝对误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大，值越接近于0，表明越相近，值越大，表明差距越大。
MaxRelativeError	最大相对误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大，值越接近于0，表明越相近，值越大，表明差距越大。
MeanAbsoluteError	平均绝对误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大。 ● MeanAbsoluteError趋于0，RootMeanSquareError趋于0，说明测量值与真实值越近似。 ● MeanAbsoluteError趋于0，RootMeanSquareError越大，说明存在局部过大的异常值。 ● MeanAbsoluteError越大，RootMeanSquareError等于或近似MeanAbsoluteError，说明整体偏差越集中。 ● MeanAbsoluteError越大，RootMeanSquareError越大于MeanAbsoluteError，说明存在整体偏差，且整体偏差分布分散。 ● 不存在以上情况的例外情况，因为RMSE ≥ MAE恒成立。
MeanRelativeError	平均相对误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大，值越接近于0，表明越相近，值越大，表明差距越大。
Mic	Microphone，麦克风。
ML	Machine Learning，机器学习。 机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。
MLL	Machine Learning Library，机器学习库。 基于opencv算子通过算法优化、neon指令等方式大幅提升opencv算子性能的机械学习库。
MTE1	Memory Transfer Engine 1，内存传输引擎1。 从L1 Buffer拷贝内存。
MTE2	Memory Transfer Engine 2，内存传输引擎2。 从DDR或者L2 Buffer拷贝内存。
MTE3	Memory Transfer Engine 3，内存传输引擎3。 从UB拷贝内存。

术语/缩略语	含义
N	
NCS	Neural Compute Server，神经计算服务器。 NCS封装AscendCL运行服务接口，可接受外部远程上板请求，并且返回对应的性能数据。
NIC	Network Interface Controller，网络接口控制器。 也称为网络接口卡，网络适配器，LAN适配器，以及类似术语。是将计算机连接到计算机网络的计算机硬件组件。
NN	Neural Network，神经网络。 在机器学习和认知科学领域，是一种模仿生物神经网络的结构和功能的数学模型或计算模型。
NPU	Neural-Network Processing Unit，神经网络处理器单元。 采用“数据驱动并行计算”的架构，特别擅长处理视频、图像类的海量多媒体业数据，专门用于处理人工智能应用中的大量计算任务。
NV	NonVolatile，永久性。 数据一旦写入NV，即使掉电也不会丢掉，下次重启，仍然会保留原有设置。
O	
OP	Operator，算子。 操作运算，比如AI的ReLU、Conv、Pooling、Scale、Softmax等。
OPAT	Operator Auto Tuning，算子调优。 OPAT是一种提升算子性能的优化器。AOE将一张整图输入给OPAT，OPAT内部进行算子融合，将融合得到的图进行算子粒度切分，针对每一个融合算子子图生成不同的算子调优策略，从而实现最优的算子性能，并将得到的最优策略保存在算子知识库。
OPP	Operator Package，算子库。
OS	Operating System，操作系统。
OTG	On-The-Go。 主要应用于各种不同的设备间的联接，进行数据交换。

P-T

表 1-4

术语/缩略语	含义
P	

术语/缩略语	含义
PCB	Printed Circuit Board，印刷电路板。 含有按预先设计形成的印制元件或印制线路以及两者结合的导电图形的印制板。
PCIe	Peripheral Component Interconnect Express，快捷外围部件互连标准。 PCIe属于高速串行点对点双通道高带宽传输，所连接的设备分配独享通道带宽，不共享总线带宽，主要支持主动电源管理，错误报告，端对端的可靠性传输，热插拔以及服务质量(QOS)等功能。
PMU	Performance Monitor Unit，性能监视单元。 PMU是CPU提供的一个单元，属于硬件的范畴。PMU通过访问相关的寄存器能读取到CPU的一些性能数据。
PNGD	PNG Decoder，PNG图像解码器。 提供对PNG格式的图像进行解码的能力。
PWM	Pulse Width Modulation，脉冲宽度调制。 脉冲载波的脉冲持续时间脉宽随调制波的样值而变的脉冲调制方式。
PyNative模式	MindSpore的动态图模式，将神经网络中的各个算子逐一下发执行，方便用户编写和调试神经网络模型。
R	
RAM	Random Access Memory，随机存储器。 基于半导体的可被CPU或者其他硬件设备读写的内存。可以任何顺序访问存储位置。
RC	Root Complex，根组件。 在PCI Express（PCIe）系统中，根组件设备将处理器和存储器子系统连接到由一个或多个交换设备组成的PCI Express交换结构。类似于PCI系统中的主机桥，根组件代表处理器生成事务请求，处理器通过本地总线互连。根组件功能可以实现为分立设备，或者可以与处理器集成。
RelativeEuclideanDistance	欧氏相对距离算法。 精度比对算法之一，计算结果取值范围为0到无穷大，值越接近于0，表明越相近，值越大，表明差距越大。
RGMII	Reduced Gigabit Media Independent Interface，精简的千兆比媒介独立接口。

术语/缩略语	含义
RootMeanSquareError	均方根误差算法。 精度比对算法之一，计算结果取值范围为0到无穷大。 ● MeanAbsoluteError趋于0，RootMeanSquareError趋于0，说明测量值与真实值越近似。 ● MeanAbsoluteError趋于0，RootMeanSquareError越大，说明存在局部过大的异常值。 ● MeanAbsoluteError越大，RootMeanSquareError等于或近似MeanAbsoluteError，说明整体偏差越集中。 ● MeanAbsoluteError越大，RootMeanSquareError越大于MeanAbsoluteError，说明存在整体偏差，且整体偏差分布分散。 ● 不存在以上情况的例外情况，因为RMSE ≥ MAE恒成立。
Runtime	Runtime运行于APP进程空间，为APP提供了针对昇腾AI处理器设备的Memory管理、Device管理、Stream管理、Event管理、Kernel执行等功能。
S	
Scalar	标量，一般表示一个常数。
SDK	Software Development Kit，软件开发工具包。 一般都是一些软件工程师为特定的软件包、软件框架、硬件平台、操作系统等建立应用软件时的开发工具的集合。
SGAT	SubGraph Auto Tuning，子图调优。 SGAT是一种提升子图性能的优化器。一张完整的网络，会被拆分成多个子图。针对每一个子图，通过SGAT生成不同的调优策略。SGAT的调优算法通过获取每个迭代的调优策略性能数据，找到最优的调优策略，从而实现对应子图的最优性能。
SoC	System on Chip，片上系统。 这是ENP成本低的关键技术，通过强大的芯片技术能力，把一个交换机单板的包转发功能全部集成到了一个芯片中，获得了软件灵活性+交换机的低成本。
SPI	Serial Peripheral Interface，串行外设接口。 SPI总线系统是一种同步串行外设接口，它可以使MCU（Microcontroller Unit，微控制单元）与各种外围设备以串行方式进行通信以交换信息。
StandardDeviation	标准差算法。 精度比对算法之一，计算结果取值范围为0到无穷大。标准差越小，离散度越小，表明越接近平均值。该列显示My Output和Ground Truth两组数据的均值和标准差，第一组展示My Output模型dump数据的数值(均值;标准差)，第二组展示Ground Truth模型dump数据的数值（均值;标准差）。
STARS	System Task and Resource Scheduler，系统任务和资源调度器。

术语/缩略语	含义
T	
TBE	Tensor Boost Engine，张量加速引擎。 提供通过Python语言实现算子的接口，能够编译生成CCE算子。
TEE	Trusted Execution Environment，可信执行环境。 在ARM Trustzone的硬件隔离环境基础上，结合硬件可信根设计，实现安全启动、安全存储、安全升级、安全运行等功能，为系统提供可信的基础运行环境。
Tensor	张量。 TensorFlow程序中的主要数据结构。张量是N维（其中N可能非常大）数据结构，最常见的是标量、向量或矩阵。张量的元素可以包含整数值、浮点值或字符串值。
Tensor比对	张量比对，两个张量之间进行不同算法评价指标的数据比对操作，支持整网比对和单算子比对。
TOPS	Trillion operations per second，每秒万亿次的运算。 用于衡量CPU、GPU、NPU的计算能力。
TS	Task Scheduler，任务调度。 通过Task Schedule分发不同的kernel到AI CPU/AI Core执行。
TVM	Tensor Virtual Machine，张量虚拟机。 TVM提供内置算子和自定义算子扩展。支持Caffe、Tensorflow等开源框架。
图模式	MindSpore的静态图模式，将神经网络模型编译成一整张图，然后下发执行。该模式利用图优化等技术提高运行性能，同时有助于规模部署和跨平台运行。

U-Z

表 1-5

术语/缩略语	含义
U	
UART	Universal Asynchronous Receiver/Transmitter，通用异步收发传输器。 用于控制计算机与串行设备的芯片。它提供了RS-232C数据终端设备接口，这样计算机就可以和调制解调器或其它使用RS-232C接口的串行设备通信。
V	

术语/缩略语	含义
VCM	Video Content Management，视频内容管理系统。 视频内容管理平台，依托于华为领先的智能图像处理及大数据分析技术，定位与视频结构化分析能力平台，提供高性能硬件、丰富的算法集成接口，有效提升图像侦查效率及业务协同能力，协助案件快速侦破，助力全球城市安全。
VDEC	Video Decoder，视频解码器。 提供对特定格式的视频进行解码的能力。
VENC	Video Encoder，视频编码器。 提供对特定格式的视频进行编码的能力。
VECTOR	向量运算
VPC	Vision Preprocessing Core，视觉预处理核心。 提供对图像进行缩放、色域转换、降bit数处理、存储格式转换、区块切割转换等能力。
Y	
YUV	Luminance-Chrominance，明亮度-带宽-色度。 Y表示明亮度（Luminance），即灰阶值，U和V表示色度（Chrominance），描述影像色彩及饱和度，用于指定像素的颜色。
Z	
整网比对	精度比对工具里Tensor比对的一种方式。对网络模型中参与计算的所有算子进行精度比对。

