ICS 国际标准分类号
CCS 中国标准文献分类号

团　　体　　标　　准

T/CES XXX-XXXX

电力人工智能算法异构硬件加速技术规范

Technical specification for heterogeneous hardware acceleration of electric power artificial intelligence algorithm

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　发布

目　　次

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分：标准化文件的结构和起草规则》给出的规则起草。

请注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别这些专利的责任。

本文件由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司、中国科学院上海微系统与信息技术研究所。

本文件主要起草人：李强、赵峰、庄莉、王秋琳、宋立华、卜智勇、王营冠、李炳森、伍臣周、何为、梁懿、陈又咏、邱镇、张晓东、李建华、陈江海、林闽微、吕志超、张维、王婧。

本文件为首次发布。

电力人工智能算法异构硬件加速技术规范

1 范围

规范规定了电力领域中人工智能算法模型训练、推理异构硬件加速的技术要求和评价方法，为电力领域中线路巡检、监控等算法模型加速提供了技术参考和评价依据。

适用于支持训练和推理的人工智能框架硬件加速技术的评估。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1-2020 标准化工作导则 第 1 部分：基本术语

GB T 41867-2022 信息技术 人工智能 术语

GB/T 5271.1-2000 信息技术 词汇 第 1 部分：基本术语

GB/T 5271.28-2001 信息技术 词汇 第 28 部分：人工智能 基本概念与专家系统

GB/T 5271.34-2006 信息技术 词汇 第 34 部分：人工智能 神经网络

T/CES 128-2022 电力人工智能平台总体架构及技术要求

YD/T 3944-2021 人工智能芯片基准测试评估方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 Artificial Intelligence

一门交叉学科，通常视为计算机科学的分支，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统。

[来源 GB/T 5271.28-2001,定义 28.01.01]

3.2

异构计算 Heterogeneous Computing

不同类型指令集合体系架构的计算单元组成系统的计算方式。

[来源：维基百科]

3.3

分布式计算 Distributed computing

是一种需要进行大量计算的工程数据分割成小块，由多台计算机机器分别计算，在上传计算结果后，将结果统一合并的得出数据结论的科学。

[来源：维基百科]

3.4

深度学习 deep learning

通过训练具有许多隐藏层的神经网络来创建丰富层次表示的方法。

[来源：GB T 41867-2022, 3.4.27]

3.5

训练 training

教会神经网络在输入值的样本和正确输出值之间做出结合的步骤。

[来源：GB/T 5271.34-2006,34.03.18]

3.6

推理 inference

从已知前提导出结论的推理方法。
注 1：在人工智能领域，前提是事实或者规则。
注 2：术语“推理”既指过程也指结果。
［来源：GB/T 5271.28-2001,28.03.01］

3.7

计算量 FLOPs

模型计算的浮点计算数，衡量模型计算的时间复杂度。

3.8

参数量 Params

模型参数所占用的字节数，衡量模型的空间复杂度。

3.9

AI 加速器 artificial intelligence accelerator

一类专用于人工智能硬件加速的微处理器或计算系统，通常由专用 AI 芯片制成，在通用或特定人工智能领域上较通用 GPU 可达到或发挥更好的性能优势。呈现形态包含但不局限于 GPU 、FPGA 、ASIC。按任务可分为训练和推理两类。

3.10

批量 batch

训练样本的一部分。
注 1：对特定计算设备，当训练样本数量过大时，可将样本分成若干批，分批训练。
注 2：批中含有的样本量是训练超参之一。
［来源：GB/T 41867-2022,3.04.21］

3.11

批次 epoch

在深度学习模型训练场景中，完整训练数据集的一次训练循环，一个 Epoch 中，模型会对整个数据集进行一次前向传播和反向传播，更新所有的参数。

3.12

迭代 iteration（in neural networks）

针对一批样本，重复地执行系列步骤直至完成训练的过程。
注 1：一个（训）期中的迭代数量等于该期中，训练样本的批数。
［来源：GB/T 41867-2022,3.04.04］

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。

AI： 人工智能（Artificial Intelligence）
GPU： 图形处理器（Graphics Processing Unit）
FPGA： 现场可程式门阵列 (Field-Programmable Gate Array)
CPU： 中央处理器（Central Processing Unit）
NPU： 神经网络处理器 (Neural-network Processing Unit)
TPU： 张量计算器 （Tensor Processing Unit）
RDMA：远程直接内存访问 （Remote Direct Memory Access）
PS： 参数服务器 (Parameter Server)
IR: 中间表示 (Intermediate Representation)
FPS: 每秒钟处理的帧数 (Frames Per Second)
QPS: 每秒钟的查询数量(Queries Per Second)

loss： 损失函数的值
MOPS： 处理器每秒钟可进行一百万次（Million Operation Per Second）
GOPS： 处理器每秒钟可进行十亿次（Giga Operations Per Second）
TOPS： 处理器每秒钟可进行一万亿次（Tera Operations Per Second）
Broadcast： 广播机制

5 电力人工智能算法异构硬件加速框架

5.1 概述

电力人工智能算法异构加速包括：训练异构硬件加速和推理异构硬件加速，其总体架构见图 1。

1）硬件加速评价指标：安装部署、模型支持与验证、训练性能测试、推理性能测试等；

2）硬件加速技术要求：分布式通信层接入接口（仅面向训练框架）、设备管理层接入接口、算子适配层接入接口要求；

3）硬件平台环境：不对硬件平台进行技术要求定义，仅规范框架适配硬件平台的环境要求。

图 1 电力人工智能算法异构硬件加速框架

5.2 电力人工智能训练异构加速（非必须）

训练流程包括数据加载 （从磁盘获取网络存储空间加载训练数据）、数据预处理 （将数据进行各种数据增强变换和尺寸处理）、前向计算 （将处理完成的数据输入网络计算 loss）、反向传播 （根据优化器，反向梯度更新，优化每一层的参数）。

训练异构加速分为单机训练模式的异构组合和多级训练模式的异构组合。

5.2.1 单机训练模式下的异构组合

单机训练模式：异构硬件在同一台物理机器上，任务间不宜进行网络通信。异构硬件工作流程如下图所示。

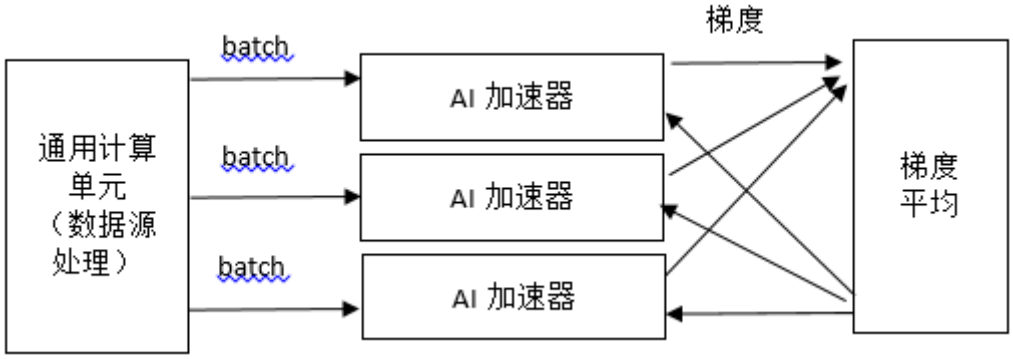

图 2 单机多卡模式工作流程

通用计算单元任务产生的数据由一个大的 batch 拆分成小的 batch 发送到 AI 加速器的内存中，每个计算单元取数据前向计算损失值 loss,反向计算梯度后需要将各个计算单元的梯度取平均值，再返回给各个计算单元更新模型参数。梯度平均值计算可以在通用计算单元或者 AI 加速器上运行。

5.2.2 多机训练模式下的异构组合

多机训练模式的异构组合包括但不限于参数服务器 PS（Parameter Server）结构和基于规约 Ring All Reduce 结构两种架构。

a）PS 结构：PS 架构的中心节点用来存储参数和梯度，由一个/一组机器组成。当更新梯度时，全局中心节点接受其他 worker 节点的数据，经参数平均法等本地计算后，再 broadcast 广播到所有其他 worker。随着 worker 数量的增加，整体通信量线性增加。

b）Ring All Reduce 结构： N（N≥2）worker 节点连接构成一个环，每个 worker 依次把自己的梯度同步给紧邻的 worker，经过至多 2*(N-1)轮同步，所有 worker 完成梯度更新。所有节点是平等的，随着 worker 的增加，整体通信量并不随着增加。

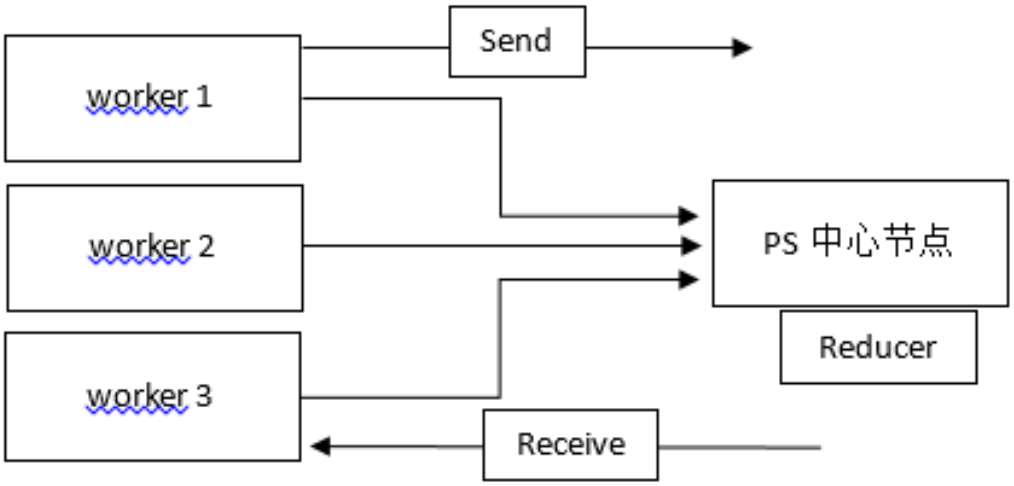

图 3 多机多卡 PS 结构

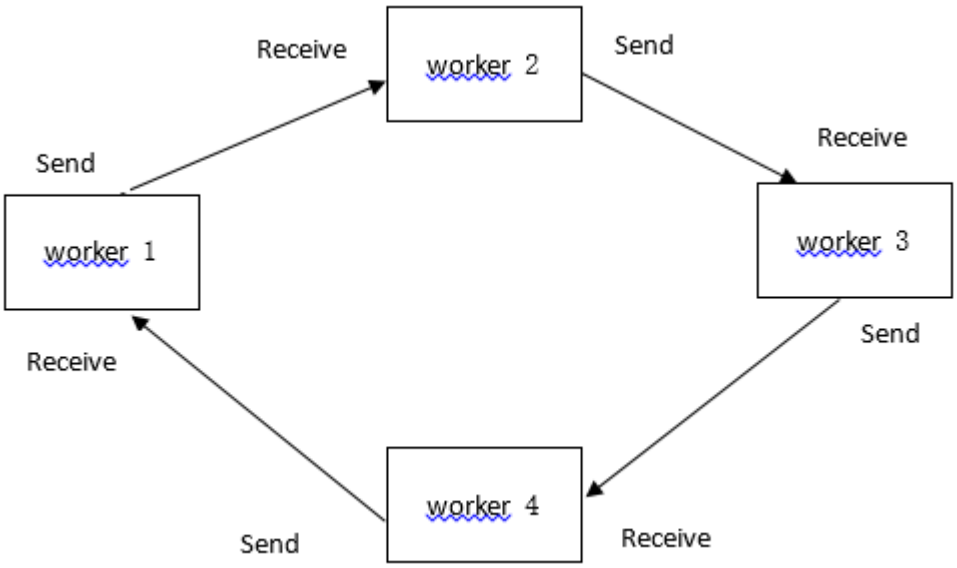

图 4 Ring All Reduce 结构

5.3 电力人工智能推理异构加速

电力人工智能推理异构加速是将训练得到的模型部署到特定异构硬件上，其流程如下图所示。

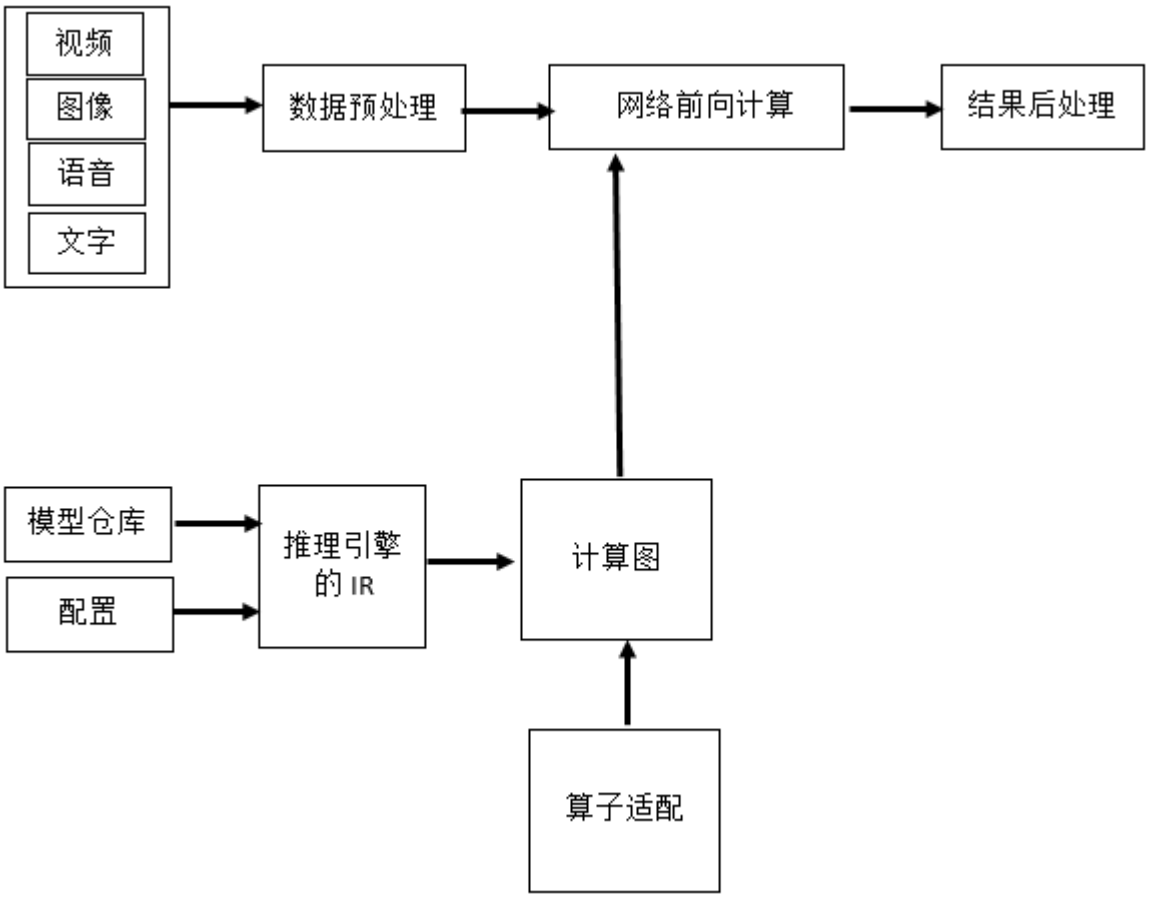

图 5 电力人工智能推理异构流程图

推理步骤如下：

a）模型 IR 转换：从模型仓库中导出训练的模型，转换工具将训练模型 IR 转换成当前异构硬件推理引擎支持的 IR。

b）计算图初始化：导入模型 IR 和当前计算单元的适配算子生成相应硬件下的计算图。

c）模型推理：数据前处理、计算图前向计算、结果后处理。

6 电力人工智能异构硬件加速的技术要求

异构硬件加速应至少包括 AI 处理器、AI 服务器、AI 集群三种之一，应符合但不限于以下要求：

a）应支持以下 1 种或多种自主可控处理器架构，自主可控处理器包括但不限于鲲鹏等；基于复杂指令集计算机架构的处理器，如 x86、x64 架构；基于精简指集计算机架构的处理器如 RISC-V、ARM、MIPS 等架构；

b) 应支持的硬件架构包括但不限于 FPGA 和 ARM 内核等；

c）应支持以下至少 1 种计算单元，包括但不限于通用处理器 CPU、GPU；ASIC 类型的加速器，如 NPU 等；FPGA 类型的加速器；GPU 类型的加速器；

d）应支持至少 1 种主流的人工智能框架，包括但不限于 TensorFlow、Pytorch、Caffe/Caffe2、Mxnet、ONNX、MindSpore（昇思）或 PaddlePaddle（飞桨）等。

e）应支持的模型精度：FP64、FP32、FP16、INT4、INT8、INT16、BP16 或混合精度等。其中，训练场景精度应支持 FP16、FP32、FP64，推理场景下精度应支持 INT8、FP16。

f) 设备管理层接口：对硬件平台驱动与运行时的接入接口进行抽象与封装，并向算子适配层、训练与推理框架提供一致的设备管理层接口。

g) 算子适配层接口：人工智能算子与目标硬件算子内核函数的映射与匹配，针对不同硬件类型规范不同的适配接口。算子层适配接口应提供算子开发或映射、子图或整图接入 2 种适配接口，宜提供编译器后端接入适配接口。硬件平台可根据环境类型的不同，选择不同的适配接口。

1) 算子开发或映射：若硬件支持可编程算子内核开发语言，或硬件具备对应的 AI 算子库，则可以选择该方式接入；

2）图引擎接入：若硬件支持图引擎，则可以选择该方式进行子图或整图接入；

3）编译器后端接入：若硬件支持编译器后端，或硬件支持代码生成器，则可以选择该方式进行人工智能编译器的算子接入。

h) 分布式通信层接口：对硬件平台的集合通信库接入框架的接口进行封装与抽象，为上层的训练框架提供一致的分布式通信层接口，允许硬件自行实现相应接口接入框架。推理框架无需实现分布式通信接口。

i）系统应考虑兼容性问题，主板接口上支持多种计算设备的接入，电源系统应能满足多种计算设备的功率需求。

6.1 电力人工智能训练异构硬件加速的技术要求（非必须）

电力人工智能训练异构硬件加速的技术要求应符合但不限于以下要求：

a）学习框架：应具备基础单卡、多卡与多机的模型训练功能；

b）操作系统：应支持基于 Linux 内核的操作系统；

c）芯片类型：应在通用 CPU 和 GPU 之外支持至少一种 AI 训练芯片；

d）设备识别：硬件驱动应支持选定操作系统的安装/卸载，设备可正确识别，宜支持容器映射；

e）人工智能算法框架应提供设备管理层接口供硬件平台的驱动和运行时接入，使硬件可被框架识别；

f）人工智能算法框架应提供硬件算子的内核函数注册接口，供目标硬件进行内核函数或相关算子库的接入；

g）应提供整图或子图组网信息与定义，由硬件平台的图引擎自行接管计算图的组网与执行并返回计算结果；

h）宜提供编译器后端接入接口规范。硬件厂商为其硬件提供编译器后端，通过编译器将框架侧的计算图模型根据特定硬件目标产生编译器端的低级 IR，然后根据硬件后端再转化为某个具体硬件上的可执行代码；

i）应提供分布式通信层接口供硬件平台的集合通信库接入，支持框架大规模分布式训练功能。

6.2 电力人工智能推理异构硬件加速的要求

电力人工智能推理异构硬件加速的技术要求应符合但不限于以下要求：

a）操作系统：宜支持 linux、windows 等常用智能终端操作系统、嵌入式操作系统等；

b）芯片类型：应在通用 CPU 和 GPU 之外支持至少一种专用 AI 推理芯片；

c）设备识别：硬件驱动应支持选定操作系统的安装/卸载，设备可正确识别，宜支持容器映射；

d）人工智能算法框架应提供设备管理层接口供硬件平台的驱动和运行时接入，使硬件可被框架识别；

e）人工智能算法框架应提供硬件算子的内核函数注册接口，供目标硬件进行内核函数或相关算子库的接入；

f）应提供子图检测和融合的能力，运行时将检测到的子图原始算子通过下发子图的方式，供硬件接管，硬件负责相关算子的调度和执行，并向框架返回输出结果；

g）宜提供编译器后端接入接口；

h）推理包括嵌入式推理和服务器推理，推理评价等级宜根据任务模型的参数量和计算量衡量，参数量、计算量都大于 0。模型的参数和参数量等级参考以下规则如表 1，表 2 所示。

表 1 模型参数大小等级

参数量(单位 MB)	级别
≥1000	C1
≥100	C2
≥10	C3
>0	C4

表 2 模型计算量大小等级

计算量（G）	级别
≥1000	C1
≥100	C2
≥10	C3
>0	C4

注：每秒操作数量 OPS（Operations per second）作为衡量硬件算力水平的一个性能指标，单位包括：

MOPS： 处理器每秒钟可进行一百万次（Million Operation Per Second）

GOPS： 处理器每秒钟可进行十亿次（Giga Operations Per Second）

TOPS： 处理器每秒钟可进行一万亿次（Tera Operations Per Second）

i）异构硬件加速部署相对于原始的训练模型输出（典型以 CPU Float32 计算为例）存在差异，差异值的均方误差作为异构硬件的精度标准，均方误差值越小，整体的推理精度越高。电力人工智能推理异构加速精度等级如表 3 所示。

表 3 推理异构硬件加速精度

输出差异均方误差	级别
< 10	C1
< 1	C2
<0.1	C3
<0.01	C4

6.2.1 电力人工智能不同场景的性能要求

电力人工智能包含有线路巡检、监控、数据分析等多种不同的应用场景，不同的应用场景对于精度、速度与存在不同的要求，场景适用等级如下表所示：

表 4 推理所需精度级别

场景	精度级别
电路巡检数据离线检测	C1
电路巡检数据实时检测	C2

配电变电监控	C3
大数据分析预测	C4

表 5 推理所需速度级别

场景	速度级别
电路巡检数据实时检测	C1
配电变电实时监控	C2
数据离线检测	C3
大数据分析预测	C4

7 电力人工智能异构加速性能评估指标及测试方法

7.1 电力人工智能模型训练异构加速性能评估指标和测试方法

7.1.1 安装部署

基于选定的基础软硬件平台，人工智能框架应具备多种安装部署能力，以便开发/测试/运维人员进行使用/管理/维护/升级等工作：

a) 应提供对应软/硬件环境下的人工智能开发框架的安装包，支持安装/卸载功能；

b）应提供对应软/硬件环境下的人工智能开发框架的容器运行镜像，支持容器内运行环境；

c）应提供对应软/硬件环境下的人工智能开发框架的容器编译镜像，支持容器内源码编译；

d) 宜支持异构 CPU 编译并支持纯 CPU 训练场景，支持 CPU 算子 kernel 优化与加速。

7.1.2 模型支持与验证

基于选定的基础软硬件平台，人工智能框架应支持在图像分类、目标检测等应用领域的人工智能模型及其评估。

7.1.3 时间

在特定数据集上训练一个模型使其达到目标准确率时的训练时间(不包括预处理和模型加载时间)。训练阶段统计的时间指标单位毫秒（ms），相关的评估指标和评估方法如下：

a) 单步训练用时

1） 定义：针对具体的训练任务，在一定大小的 batch 输入，进行一次前向传播反向梯度更新的计算过程的耗时。

2） 测量方法：

--在 batch 数据送入 input 节点的时间记为 T_0；

--在梯度更新完成的时间记为 T_1；

--单步训练的用时就为 $T_1 - T_0$。

说明：单个 step 的时间可以描述异构计算中的纯粹网络计算的时间和梯度更新时间，这个时间越短就越好，同时单个 step 下可以针对不同的 batch 的维度进行时间的统计，在时间比较上采用相同的 batch 下进行比较。

b) 总的训练用时

1）定义：针对一个具体的训练任务，在训练数据一定，epoch 一定时，总的任务所花费的时间。

2）测量方法：

--启动训练的脚本的时间设置为 Ts；

--训练程序退出的时间设置为 Te；

--总的训练时间为 Te - Ts。

说明：总的训练时间包含模型的初始化，数据加载，网络训练计算，梯度更新模型保存等所有的时间，这个可以反映整个人工智能算法训练异构加速系统的整体性能，时间越短性能越高。

7.1.4 实际计算利用率

训练阶段统计异构计算中计算设备在一段训练时间内的实际利用率。

a) 计算设备的使用率

1）定义：统计计算设备在一个训练周期内的计算使用率。

2）方法：

--在一个 epoch 的训练过程中间隔 1s 采样获取计算设备的使用率 Si；

--统计一个 epoch 中所有的使用率的总和 S 以及采样的个数 N；

--单个 epoch 中计算单元的使用率为 S/N。

说明：计算设备的利用率反映了整个训练系统在异构硬件加速上的使用效率，这个指标越高说明系统采用设备的硬件加速越明显，整体的效果越明显。

7.1.5 吞吐率

吞吐率反应了整个异构硬件加速系统针对训练业务的计算能力，单位是 MB/s。

a) 单个节点异构硬件的吞吐率

1）定义：在训练过程中单 EPOC 时间内处理的数据量和时间的比值。

2）方法：

--统计单个 epoch 的训练处理时间 T_i；

--统计各个 epoch 的平均训练处理时间 T_a；

--最终就是一个 epoch 的训练样本数量/T_a。

说明：实际的吞吐率反映了单台机器上异构硬件针对训练过程中的数据加载、数据预处理、网络前向计算、反向传播更新梯度全流程的能力。

b) 集群系统异构硬件的吞吐率

1）定义：同上

2）方法：

-- 统计单个节点上异构硬件系统的吞吐率 Thi；

-- 所有节点上的 Thi 进行平均就是系统的 The。

7.1.6 功耗

功耗是以瓦（W）为单位，反映了异构设备在实际训练中功率使用情况。

a) 平均功耗

1）定义：在整个训练过程中的平均功率。

2）测量方法：

--使用功率计周期采样测量整机的功率；

--求取平均值就是训练过程的平均功耗。

说明：平均功耗反映了整个异构加速系统在训练中的能源使用情况。

b) 峰值功率

1）定义：在训练过程中异构设备瞬时最大的功率。

2）测量方法：

--通过提高 batch 提高整个异构设备的负载压力；

--用功率计采集满负荷下的运行功率，选择功率做大的一个值作为峰值功率。

说明：峰值功率反映了异构加速系统在使用过程中的最大功率。

7.1.7 能效

能效指的是在单位能耗下训练处理的数量的多少，单位采用（MB/（W*S））。

a) 单机能效

1）定义：单台机器上异构设备在单位能耗下训练处理的数量的多少。

2）方法：

-- 统计每个 epoch 运行期间设备的用电量 Ei；

-- 对所有的 epoch 期间的电量求平均 E；

-- 可以得到最终的能效为 N/E（这里的 N 是一个 epoch 的数据数量）。

b) 集群能效

1） 定义： 集群在单位能耗下训练处理的数据的数量多少。

2）方法：

-- 统计每个 epoch 运行期间集群的用电量 Ei；

-- 对所有的 epoch 期间的电量求平均 E；

-- 可以得到最终的能效为 N/E （这里的 N 是一个 epoch 的数据数量）。

能效也可以反映出异构硬件在加速算法训练过程中的能源利用情况，能效越高整个异构加速硬件的能力越高。

7.2 电力人工智能模型推理异构加速性能评估指标和测试方法

7.2.1 安装部署

基于选定的基础软硬件平台，人工智能框架应应具备多种安装部署能力，以便开发/测试/运维人员进行使用/管理/维护/升级等工作：

a) 应提供对应软/硬件环境下的人工智能推理框架的安装包，支持安装/卸载功能；
b) 应提供对应软/硬件环境下的人工智能推理框架的 C/C++推理库，支持模型部署上线；
c) 应提供对应软/硬件环境下的人工智能推理框架的容器运行镜像，支持容器内运行环境；
d) 应提供对应软/硬件环境下的人工智能推理框架的容器编译镜像，支持容器内源码编译。

7.2.2 模型支持与验证

基于选定的基础软硬件平台，深度学习框架应支持基础模型，结果正确，性能符合对应硬件预期。

7.2.3 时间

推理阶段统计的时间指标单位毫秒（ms），相关的评估指标和评估方法如下：

a） 单个数据的平均推理时间

1） 定义：batch 大小为 1 的数据完成数据前处理、数据拷贝到计算单元、计算单元网络前向传播、从计算单元拷贝出来、结果后处理的总时间。

2） 测量方法：

--将整个数据前处理、数据拷贝到计算单元、计算单元网络前向传播、从计算单元拷贝出来、结果后处理封装成一个模块，将 N 个测试数据分别传入模块中去计算，获取每个数据调用模块的耗时 Ti。

3） 计算最终的平均推理时间为$\frac{1}{N}\sum_{0}^{N-1}Ti$。

针对嵌入式实时场景，在功耗、输入数据相同的情况下，单个数据的平均处理时间越短，整个异构硬件加速更好。

7.2.4 FPS

FPS 反应了整个异构硬件加速系统针对推理业务的计算能力，单位是 MB/s。

a） 单个计算节点的 FPS

1）定义：单位时间内，单个计算节点处理的数据的数量。

2）测量方法：

--选取 N 个测试的数据；

--统计每个数据经过推理模块的耗时 Ti；

--将 N 个时间相加得到 Ts；

--最终的 FPS 就是为 N/Ts。

b） 计算集群的 FPS

1）定义：单位时间内，计算集群处理的数据的数量。

2）测量方法：

--选取 N 个测试数据；

--将 N 个测试数据平均分配到 M 个计算节点上；

--统计每个节点上计算任务的开始时间 Tis 和介绍时间 Tio；

--从 M 个 Tis 中找到最小的时间 Tismin；

--从 M 个 Tio 中找到最大的时间 Tiomax；

--整个计算系统的总耗时为 T= Tiomax - Tismin；

--整个计算集群的 FPS 为 N/T。

针对嵌入式实时场景，在网络模型一定、数据一定的情况下，这个参数越大，反应异构加速能力越强。

7.2.5 QPS

QPS 反映出异构硬件服务器的推理服务提供能力，单位是 MB/s。

a） 单个服务器的最大 QPS

1）定义：在给定的响应时延范围内，单个异构服务器单位时间最大的处理次数。

2）测量方法：

--客户端安装 jmeter 压测工具；

--jmeter 设定平均的响应时间；

--客户端会根据平均响应时间设置不同的请求线程数进行压测，jemter 获取对应的 Qi；

--选择 Qi 最大的值作为最大的 QPS。

b) 服务器集群的最大 QPS

1）定义：在给定的响应时延范围内，异构服务器集群单位时间最大的处理次数。

2）测量方法：

--跟单机时测量方式一样。

说明：最大 QPS 反映出了，在服务器端推理场景下的异构加速服务器的处理能力，这个值越大越好。

7.2.6 计算资源的利用率

在推理阶段异构体系中计算资源的实际利用率。

a) 单个计算设备的最大利用率

1）定义：在满负荷的条件下整个推理过程中单个计算设备的平均资源利用率。

2）测量方法：

--编写定时采集资源利用率的脚本；

--设置好最大的时延时间；

--在给定的时延下，加大服务请求数量或者直接加大 batch 和多流的处理方式，使系统的负载达到最大；

--运行脚本定时采样资源利用率；

--对采样的数据求取平均，即单个计算设备在满负荷推理的最大设备利用率。

b) 异构计算集群计算设备的最大利用率

1）定义：在满负荷的条件下整个推理过程中集群中计算设备的平均资源利用率。

2）测量方法：

--编写定时采集资源利用率的脚本部署在集群各个节点；

--设置整个集群计算的最大响应时延时间；

--按照上面的测试方法测试各个计算设备的最大利用率 Si；

--针对各个节点的 Si 进行求取平均就是异构计算集群的计算资源的利用率。

说明：针对特定的算法模型，计算资源利用率反映出在满足特定时延下，可以发挥出的计算能力多大，这个越大，说明系统计算资源使用越多，整体加速越明显。

7.2.7 功耗

功耗是以瓦（W）为单位，反映了异构设备在实际推理中功率使用情况。

a)平均功耗

1）定义：在整个推理过程中的平均功率

2）测试方法：

--配置使用功率计；

--在给定时延范围内，周期测量整机的功率；

--对功率求取平均值。

说明：平均功耗反映了整个异构加速系统在推理中的能源使用情况。

b)峰值功率

1）定义：服务器进行压力测试下，满负荷时的最大瞬时功率

2）测试方法：

--配置功率计；

--在给定的时延的条件下，加大服务请求数量或者直接加大 batch 和多流的处理方式，使系统的负载达到最大。周期性测量功率；

--从采样数据中选取最大的值作为异构加速系统的峰值功率。

说明：峰值功率反映了异构推理系统在使用过程中的最大功率。

7.2.8 能效

能效指的是在单位能耗下推理处理的数量的多少，单位采用（MB/（W*S））。

a)单机能效

1）定义：单个异构机器在单位能耗下，能处理的数据量的多少。

2）测量方法：

--选择 N 个数据组成的推理测试数据集；

--设置单个请求的最大的时延 T；

--在给定的时延 T 下，N 个数据采用多任务、多 batch、多流的方式

在当前的单机异构硬件环境下进行处理；

--通过设备测量当前机器在处理完成这些数据所需要的电量 E；

--最终的能效为 N/E。

b) 集群的能效

1）定义：在异构机器集群在单位能耗下，能处理的数据量的多少。

2）测量方法：

--方法同单机，电量统计时需要将各个计算节点的电量求和得到 Es，集群的能效是 N/Es。

参 考 文 献

[1] T/CES 128-2022 电力人工智能平台总体架构及技术要求.

[2] YD/T 3944-2021 人工智能芯片基准测试评估方法.

[3] ISO/IEC FDIS 22989-2022 Information technology Artificial intelligence-Artificial intelligence concepts and terminology[S].