

人工智能 加速卡技术要求及测试方法

1 范围

本文件规定了人工智能加速卡的技术要求，包括人工智能加速卡的通用技术要求、人工智能训练卡和推理卡的功能要求、性能要求、兼容性要求、可靠性要求、性能度量指标，以及人工智能加速卡的安全性要求，并给出了人工智能训练卡和推理卡的测试环境及测试方法。本文件规定的技术要求主要面向于数据中心或服务器使用的人工智能加速卡。

本文件适用于人工智能加速卡的生产方、评测方、使用方等对加速卡进行设计、测试、评估、选型和应用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 9813.3-2017 计算机通用规范 第3部分：服务器

GB/T 17235.1-1998 信息技术 连续色调静态图像的数字压缩及编码 第1部分：要求和指南

GB/T 34986-2017 产品加速试验方法

GB/T 37092-2018 信息安全技术 密码模块安全要求

GB/T 41867-2022 信息技术 人工智能 术语

GB/T 42018-2022 信息技术 人工智能 平台计算资源规范

GM/T 0008-2012 安全芯片密码检测准则

GM/T 0028-2014 密码模块安全技术要求

YD/T 3944-2021 人工智能芯片基准测试评估方法

YD/T 4398-2023 电信行业云原生平台架构与技术要求

T/CESA 1303-2023 人工智能 异构人工智能加速器统一接口

ISO/IEC 15948:2004 信息技术 计算机图形和图像处理 便携式网络图形：功能规范 [Information technology – Computer graphics and image processing – Portable Network Graphics (PNG) : Functional specification]

ITU-T H.264—2021 通用视听服务高级视频编码（Advanced video coding for generic audiovisual services）

ITU-T H.265—2021 高效视频编码（High efficiency video coding）

3 术语和定义

GB/T 41867-2022、GB/T 42018-2022和YD/T 3944-2021界定的以及下列术语和定义适用于本文件。为了便于使用，以下重复列出了GB/T 41867-2022、GB/T 42018-2022和YD/T 3944-2021中的某些术语和定义。

3.1

人工智能系统 artificial intelligence system

针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。

[来源：GB/T 41867-2022，3.1.8]

3.2

人工智能 artificial intelligence

人工智能系统（3.1）相关机制和应用的研究和开发。

[来源：GB/T 41867-2022，3.1.2]

3.3

人工智能服务器 artificial intelligence server

信息系统中能够为人工智能应用提供高效能计算处理能力的服务器。

注 1：以通用服务器为基础，配备人工智能加速卡后，为人工智能应用提供专用加速能力的服务器，称人工智能兼容服务器。

注 2：专为人工智能加速计算设计，提供人工智能专用计算能力的服务器，称人工智能一体机服务器。

[来源：GB/T 41867-2022，3.1.3]

3.4

人工智能集群 artificial intelligence cluster

遵循统一控制的，人工智能计算功能单元的集合。

注 1：人工智能计算功能单元可包含人工智能加速处理器、人工智能服务器、人工智能加速模组等。

注 2：当由人工智能服务器组成时，人工智能集群可称为人工智能服务器集群，其中的人工智能服务器可称为节点。

[来源：GB/T 41867-2022，3.1.4]

3.5

人工智能加速卡 artificial intelligence accelerating card

专为人工智能计算设计、符合人工智能服务器硬件接口的扩展加速设备，简称“加速卡”。

[来源：GB/T 42018-2022，3.6]

注：文中的人工智能加速卡主要面向数据中心或服务器场景使用。

3.6

人工智能训练加速卡 artificial intelligence training accelerating card

一种旨在加快人工智能模型训练过程的集成电路板卡，简称“训练卡”。具备高度并行处理能力，可以显著提升人工智能模型的训练速度和效率。

3.7

人工智能推理加速卡 artificial intelligence inference accelerating card

一种旨在加快人工智能模型推理过程的集成电路板卡，简称“推理卡”。主要用于部署已经训练好的模型并在实际应用中推断和预测。推理卡通常具备高效的计算能力和低延迟，可提升人工智能模型在实时应用中的性能和响应速度。

3.8

批次大小 Batch size

单次处理时输入的样本（如图像，时间序列等）数量。

[来源：YD/T 3944-2021，3.1.9]

3.9

云原生 cloud native

是面向云应用设计的一种思想理念，充分发挥云效能的最佳实践路径，帮助企业构建弹性可靠、松耦合、易管理、可观测的应用系统，提升交付效率，降低运维复杂度。

[来源：YD/T 4398-2023, 3.1]

4 缩略语

ARM: 高级精简指令集处理器 (Advanced Reduced Instruction Set Computer Machines)

BF16: 脑浮点数 (Brain Floating Point)

CPU: 中央处理单元 (Central Processing Unit)

ECC: 错误纠正码 (Error Correcting Code)

FP8: 8位浮点数 (8-bit Floating Point)

FP16: 半精度浮点数 (Half-precision Floating Point)

FP32: 单精度浮点数 (Single-precision Floating Point)

FPS: 帧每秒 (Frames Per Second)

HBM: 高带宽内存 (High Bandwidth Memory)

INT8: 8位四分之一精度整型 (8 bits quarter-precision INTeget)

LPDDR: 低功率双数据率 (Low Power Double Data Rate)

MTBF: 平均无故障工作时间 (Mean Time Between Failure)

OAM: 开放计算项目加速器模块 (Open Compute Project Accelerator Module)

PCIe: 外设部件互联高速通道 (Peripheral Component Interconnect Express)

RDMA: 远程直接内存访问 (Remote Direct Memory Access)

RoCE: 以太网上的远程直接内存访问 (RDMA over Converged Ethernet)

TF32: 张量单精度浮点数 (Tensor Float-32)

TFLOPs: 一万亿次浮点运算 (Tera Floating-point Operations)

TFLOPS: 每秒一万亿次浮点运算 (Tera Floating-point Operations Per Second)

TOPS: 每秒一万亿次运算 (Tera Operations Per Second)

TRNG: 真随机数发生器 (True Random Number Generator)

5 概述

人工智能加速卡是一种硬件设备，具有适配人工智能算法运算微架构、能够完成人工智能应用运算处理的集成电路元件，可作为服务器的一部分，与其他组件（如CPU、存储设备等）协同工作，提供更高效、更快速的计算能力。人工智能加速卡通常由多个处理器和内存组成，这些处理器能够高效地进行矩阵计算，使其相比于CPU具有更加出色的计算能力和效率，可加速人工智能模型的训练和推理速度。

人工智能加速卡被广泛应用于人工智能领域，如计算机视觉、自然语言处理、语音识别等场景。人工智能加速卡根据不同场景下对其功能和性能等的要求不同，通常可分为人工智能训练加速卡（简称训练卡）和人工智能推理加速卡（简称推理卡）。本文件规定的人工智能加速卡主要为面向数据中心及云端的产品。

本文件主要技术内容分为三个部分，包括技术要求、测试环境和测试方法。技术要求部分给出人工智能加速卡的通用技术要求，训练卡和推理卡的功能要求、性能要求、兼容性要求、可靠性要求、训练/推理性能度量指标，以及安全性要求。测试环境部分针对技术要求中的各项内容给出测试所需明确的

测试对象、测试组网、系统配置和环境条件。测试方法部分针对技术要求部分提出的各项要求，分别给出相应的测试方法。

6 技术要求

6.1 加速卡通用技术要求

人工智能加速卡的通用技术要求如下：

- a) 应内置生产厂家、产品型号、序列号、固件版本、显存信息等基础配置信息，且可被配置了加速卡的设备（如服务器）操作系统正常读取；
- b) 应支持资产管理功能，可通过服务器远程管理系统读取加速卡的序列号信息，且该信息应与配置了加速卡的设备操作系统读取的信息保持一致；
- c) 应支持固件版本管理功能，可通过服务器远程管理系统读取加速卡的固件版本信息，且该信息应与配置了加速卡的设备操作系统读取的信息保持一致；
- d) 应支持功耗监控功能，可通过服务器远程管理系统读取加速卡的当前功耗信息，且该信息与配置了加速卡的设备操作系统读取的功耗值差距应在 5%以内；
- e) 应支持温度监控功能，可通过服务器远程管理系统读取加速卡的当前温度信息，且该信息与配置了加速卡的设备操作系统读取的温度值差距应在 5%以内；
- f) 应支持至少一种 Linux 操作系统，操作系统版本可参考附录 A.1；
- g) 应支持至少一种满足信息技术应用创新要求的操作系统；
- h) 应具备与 CPU 的卡间通信功能；
- i) 应支持加速卡性能分析工具；
- j) 宜具备虚拟化功能，支持通过虚拟化软件对整张物理加速卡进行切分；
- k) 应具备电流过载或功率过载的保护机制；
- l) 应具备面向业务负载的动态功耗性能管理机制；
- m) 应具备针对错误或异常的处理及上报机制；
- n) 应具备云原生的接入能力，如支持 K8s 等技术。

6.2 训练卡技术要求

6.2.1 训练卡功能要求

训练卡的功能要求如下：

- a) 应支持 BF16、FP16、FP32 数据精度类型；
- b) 宜支持 INT8、FP8、TF32 数据精度类型中的一种或多种；
- c) 应支持混合精度训练；
- d) 应支持自定义算子开发功能，如矩阵乘法、卷积等；
- e) 宜支持 T/CESA 1303-2023 中给出的算子类型；
- f) 应支持服务器内部训练卡的卡间高速互联通信功能；
- g) 应支持服务器之间训练卡间的高性能通信能力（如支持 RoCE、InfiniBand 等 RDMA 技术）；
- h) 应支持集合通讯库及典型的集合通信算法（如 all reduce、all gather 等），具备卡间集合通信能力；
- i) 应支持数据并行、流水线并行、张量并行等并行策略中的一种或多种；
- j) 宜支持软件加速库，通过软件层面优化加速模型训练；

- k) 宜直接具备或与解码器配合实现图像和视频的解码能力（支持 ITU-T H.264-2021、ITU-T H.265-2021 等规定的视频格式中的一种或多种,支持 GB/T 17235.1-1998、ISO/IEC 15948:2004 等规定的图像格式中的一种或多种）。

6.2.2 训练卡性能要求

训练卡的性能要求如下：

- a) 峰值计算性能宜不低于 200 TOPS (INT8)、96 TFLOPS (BF16)、96 TFLOPS (FP16)、24 TFLOPS (FP32)；
- b) 显存容量宜不小于 32GB；
- c) 面向大模型等训练场景，显存带宽宜不低于 600GB/s；
- d) 面向大模型等训练场景，节点内卡间互联聚合带宽宜不低于 200GB/s（双向）。

6.2.3 训练卡兼容性要求

训练卡的兼容性要求如下：

- a) 应支持至少一种深度学习框架；
- b) 应支持至少一种分布式训练框架；
- c) 应支持 PCIe 接口或 OAM 接口；
- d) PCIe 接口形态卡应支持 PCIe4.0 或以上版本接口协议中的至少一种；
- e) OAM 接口形态卡应支持 OAM1.1 或以上版本接口协议中的至少一种；
- f) 应支持 HBM、GDDR、LPDDR 等高带宽内存中的一种或多种。

6.2.4 训练卡可靠性要求

训练卡的可靠性要求如下：

- a) 应支持模型断点续训功能，能够自动断点保存、故障诊断与上报、自动恢复训练等；
- b) 应通过 3×24 小时压力测试；
- c) 宜支持内存错误修复（如基于 ECC）；
- d) 宜具备在受控环境中的理想条件及非受控环境中的各种环境压力条件下的 MTBF 测试结果。

6.2.5 训练性能度量指标

6.2.5.1 训练时间

训练时间是指在特定数据集上训练一个模型使其达到目标准确率时所用的时间（不包括预处理和模型加载时间），一般采取运行多次去掉最低和最高的数字后取平均值。对于大规模预训练模型，可使用模型在特定数据集上训练一轮或多轮所用的时间来衡量。训练时间的测量方法如表1所示。

表1 训练时间测量方法

度量指标	说明	测量方法
训练时间	从训练开始命令调用到训练退出之间的时间间隔	<ul style="list-style-type: none"> a) 训练开始前，串行并紧邻调用计时命令，获得时间点t_1； b) 训练退出时，串行并紧邻调用计时命令，获得时间点t_2； c) 计算训练用时：$T = t_2 - t_1$。

6.2.5.2 训练吞吐率

训练吞吐率体现了训练卡对选定的模型训练任务的计算能力。对视觉类测试，单位为图片数每秒（images/s）；对自然语言处理类测试，单位为句数每秒（sentences/s）；对自然语言语句生成模型，吞吐率为定长输入（句中单词或字的个数）、输出条件下，每秒处理的语素数量，单位是tokens/s。训练吞吐率的测量方法如表2所示。

表2 训练吞吐率测量方法

度量指标	说明	测量方法
训练吞吐率	训练卡在训练过程中，每个训期处理的数据量与时间的比值。	<p>a) 统计每个训期<i>i</i>所使用的时间T_i，计算每训期平均时间T；</p> <p>b) 训练吞吐率计算公式为：</p> $\text{训练吞吐率} = \frac{\text{number of (训练集)}}{T}$ <p>c) 对文本生成类的训练任务，训练吞吐率为：</p> $\text{训练吞吐率} = \frac{\text{number of tokens (训练集)}}{T}$ <p>其中： number of (*)表示计量特定数据集合所含的样本数量； $\text{number of tokens (*)}$表示计量特定数据集合所含的语素数量。</p>

6.2.5.3 训练功耗

训练功耗是指训练卡在执行模型训练任务期间，单位时间内所消耗的能源的值，单位为瓦（W）。训练功耗的测量方法如表3所示。

表3 训练功耗测量方法

度量指标	说明	测量方法
训练功耗	在执行训练任务期间，周期性测量被测设备的负载功率，并计算均值。	<p>计算公式为：</p> $P = \frac{1}{N} \sum_{i=1}^N P_i$ <p>其中： P：有效期内的平均输入功率； P_i：有效期内得到的输入功率值为$\{P_1, P_2, \dots, P_N\}$； N：次数。</p>

6.2.5.4 训练能效

训练能效是指训练卡在模型训练过程中，针对特定的数据精度，单位时间内消耗单位功耗，所完成的计算量，单位为万亿次浮点运算次数每瓦（TFLOPs/W）。训练能效的测试方法如表4所示。

表4 训练能效测量方法

度量指标	说明	测量方法
训练能效	训练卡单位时间内消耗单位功耗所完成的计算量。	计算公式为： $E = \frac{\text{size of (训练集)}}{T \times P}$ 其中： <i>size of (*)</i> : 针对训练集，计算并转化为浮点运算次数，单位是TFLOPs； <i>T</i> : 每个训期的平均用时； <i>P</i> : 训练任务中每个训期的平均功率，可参照6.2.5.3的方法计算。

6.2.5.5 全精度训练能效

全精度训练能效是指训练卡在模型训练过程中，针对所支持的全部数据精度的训练能效总和。全精度训练能效的测量方法如表5所示。

表5 全精度训练能效测量方法

度量指标	说明	测量方法
全精度训练能效	训练卡所支持的全部数据精度的训练能效总和，分为张量算效和矢量算效。	计算公式为： $\text{张量算效} = \sum_{i=1}^n \frac{T_i}{W_i}$ $\text{矢量算效} = \sum_{i=1}^n \frac{V_i}{W_i}$ 其中： <i>n</i> : 支持的数据精度种类数量； <i>T_i</i> : 在第 <i>i</i> 种数据精度下进行模型训练的张量峰值算力，单位为GOPS； <i>V_i</i> : 在第 <i>i</i> 种数据精度下进行模型训练的矢量峰值算力，单位为GFLOPS； <i>W_i</i> : 在第 <i>i</i> 种数据精度下进行模型训练时的平均功耗。

6.2.5.6 多卡训练线性度

多卡训练线性度用来衡量加速卡在模型集群规模化训练下的可扩展性，可分为卡线性度和集群线性度。多卡训练线性度的测试方法如表6所示。

表6 多卡训练线性度测量方法

度量指标	说明	测量方法
卡线性度	选取某一模型训练任务，计算使用单台服务器节点的多张卡并行训练时每秒处理的样本数量，与使用单张卡训练时每秒处理的样本数量之间的比值，再用此比值除以卡的数量。	计算公式为： $\text{卡线性度} = \frac{V_1}{N * V_2}$ 其中： <i>V₁</i> : 采用 <i>N</i> 张卡并行训练时每秒处理的样本数量； <i>V₂</i> : 采用单张卡训练时每秒处理的样本数量。
集群线性度	选取某一模型训练任务，计算	计算公式为：

	<p>使用包含多台服务器节点及多张训练卡的集群开展并行训练时每秒处理的样本数量，与使用单张卡训练时每秒处理的样本数量之间的比值，再用此比值除以集群中卡的数量。</p>	$\text{集群线性度} = \frac{V_3}{N * V_4}$ <p>其中： V_3: 采用包含N张卡的集群训练时每秒处理的样本数量； V_4: 采用单张卡训练时每秒处理的样本数量。</p>
--	---	--

6.2.5.7 检查点保存和加载时间

检查点保存时间是指在模型训练过程中的任一时间点，将模型训练状态保存到存储设备所需要的时间（包括模型参数和优化器等模型状态的存储时间）；检查点加载时间是指将存储设备中的模型状态（包括模型参数和优化器等模型状态）加载到训练卡上并开始模型正常训练所需要的时间，一般采用运行多次去掉最低和最高的数字后取平均值。检查点保存和加载时间的测量方法如表7所示。

表7 检查点保存和加载时间测量方法

度量指标	说明	测量方法
保存时间	将模型状态从训练卡保存到存储设备所需的时间	a) 模型状态保存开始前，串行并紧邻调用计时命令，获得时间点 t_1 ； b) 模型状态保存完成后，串行并紧邻调用计时命令，获得时间点 t_2 ； c) 计算保存用时： $T_1 = t_2 - t_1$ 。
加载时间	从存储设备中将模型状态加载到训练卡中所需的时间	a) 模型状态加载开始前，串行并紧邻调用计时命令，获得时间点 t_3 ； b) 模型状态加载完成后，串行并紧邻调用计时命令，获得时间点 t_4 ； c) 计算加载用时： $T_2 = t_4 - t_3$ 。

6.3 推理卡技术要求

6.3.1 推理卡功能要求

推理卡的功能要求如下：

- a) 应支持 INT8、FP16 数据精度类型；
- b) 宜支持 FP8、BF16、FP32、TF32 数据精度类型中的一种或多种；
- c) 应支持自定义算子开发功能，如矩阵乘法、卷积等；
- d) 宜支持 T/CESA 1303-2023 中给出的算子类型；
- e) 宜支持服务器内部推理卡的卡间高速互联通信功能；
- f) 宜支持服务器之间推理卡间的高性能通信能力（如支持 RoCE、InfiniBand 等 RDMA 技术）；
- g) 应支持集合通讯库及典型的集合通信算法（如 all reduce、all gather 等），具备卡间集合通信能力；
- h) 应支持通过模型推理服务部署工具实现模型的在线服务部署和运行；
- i) 应支持流水线并行、张量并行等并行策略中的一种或多种；
- j) 宜直接具备或与解码器配合实现图像和视频解码能力（支持 ITU-T H. 264-2021、ITU-T H. 265-2021 等规定的视频格式中的一种或多种，支持 GB/T 17235.1-1998、ISO/IEC 15948:2004 等规定的图像格式中的一种或多种）。

6.3.2 推理卡性能要求

推理卡的性能要求如下：

- a) 峰值计算性能宜不低于 200 TOPS(INT8) 、96 TFLOPS(BF16) 、96 TFLOPS(FP16) 、24 TFLOPS(FP32)；
- b) 显存容量宜不小于 16GB；
- c) 面向大模型等训练场景，显存带宽宜不低于 300GB/s；
- d) 宜支持不低于 96 路 H. 264、H. 265 格式视频解码（1080P@30FPS）；
- e) 宜支持不低于 4000 Frames/s 分辨率为 1080P 的 JPEG 格式图片解码。

6.3.3 推理卡兼容性要求

推理卡的兼容性要求如下：

- a) 宜支持至少一种深度学习框架；
- b) 应支持至少一种深度学习推理引擎；
- c) 应支持 PCIe 接口或 OAM 接口；
- d) PCIe 接口形态卡应支持 PCIe4.0 或以上版本接口协议中的至少一种；
- e) OAM 接口形态卡应支持 OAM1.1 或以上版本接口协议中的至少一种；
- f) 应支持 HBM、GDDR、LPDDR 等高带宽内存中的一种或多种；
- g) 应支持部署运行经过模型格式转化的由不同厂商生产的加速卡所训练生成的模型。

6.3.4 推理卡可靠性要求

- a) 推理过程稳定运行，同一个模型连续多次推理的推理时延波动应相当；
- b) 应通过 3×24 小时压力测试；
- c) 宜支持内存错误修复（如基于 ECC）；
- d) 宜具备在受控环境中的理想条件及非受控环境中的各种环境压力条件下的 MTBF 测试结果。

6.3.5 推理性能度量指标

6.3.5.1 推理时延

推理时延是指推理任务从执行到结束的运行时间，即从内存发送样本数据到模型输出推理结果的时间间隔。对于自然语言语句生成类任务，使用首字时延和模型稳定输出时单token的生成时间来衡量推理时延性能。推理时延的测试方法如表8所示。

表8 推理时延测量方法

度量指标	说明	测量方法
推理时延	非自然语言语句生成类任务，计算被测设备对某样本推理的开始时间与结束时间的时间间隔	a) 推理开始前，串行并紧邻调用计时命令，获得时间点 t_1 ； b) 推理退出时，串行并紧邻调用计时命令，获得时间点 t_2 ； c) 计算推理用时： $T_1 = t_2 - t_1$ 。
	自然语言语句生成类任务的首字时延，为被测设备对某样本推理的开始时间与模型输出首个token的时间间隔	a) 推理开始前，串行并紧邻调用计时命令，获得时间点 t_3 ； b) 输出首个token时，串行并紧邻调用计时命令，获得时间点 t_4 ； c) 计算推理用时： $T_2 = t_4 - t_3$ 。

	自然语言语句生成类任务的单token生成时间，为被测设备稳定生成每个token所用的时间	a) 模型稳定输出时，选定某一token，在token生成后，串行并紧邻调用计时命令，获得时间点 t_5 ； b) 当模型生成下一个新的token后，串行并紧邻调用计时命令，获得时间点 t_6 ； c) 计算推理用时： $T_3 = t_6 - t_5$ 。
--	--	---

6.3.5.2 推理吞吐率

推理吞吐率代表了推理卡对特定推理任务的计算能力。对视觉类测试，单位为图片数每秒（images/s）；对自然语言处理类测试，单位为句数每秒（sentences/s）；对自然语言语句生成的模型，吞吐率为定长输入（句中单词或字的个数）、输出条件下，每秒处理的语素数量，单位是tokens/s。推理吞吐率的测试方法如表9所示。

表9 推理吞吐率测量方法

度量指标	说明	测量方法
推理吞吐率	推理卡在单位时间内，对于特定任务负载，完成处理的样本数量。	a) 计算整个推理测试过程的时间 T ； b) 推理吞吐率计算公式为： $\text{推理吞吐率} = \frac{\text{number of (推理集)}}{T}$ c) 对文本生成类的推理任务，推理吞吐率为： $\text{推理吞吐率} = \frac{\text{number of tokens (推理集)}}{T}$ 其中： <i>number of (*)</i> ：整个推理测试过程中，由所有实际发送的样本以及实际返回结果所计算得出的样本数量； <i>number of tokens (*)</i> ：统计整个推理测试过程中，由所有实际发送的样本以及实际返回结果所计算得出的样本数量，再针对每个样本累计语素数量。

6.3.5.3 推理功耗

推理功耗是指推理卡在执行模型推理任务期间，单位时间内所消耗的能源的值，单位为瓦（W）。推理功耗的测量方法如表10所示。

表10 推理功耗测量方法

度量指标	说明	测量方法
推理功耗	在执行推理任务期间，周期性测量被测设备的负载功率，并计算均值。	计算公式为： $P = \frac{1}{N} \sum_{i=1}^N P_i$ 其中： P ：有效期内的平均输入功率； P_i ：有效期内得到的输入功率值为 $\{P_1, P_2, \dots, P_N\}$ ；

		N : 次数。
--	--	-----------

6.3.5.4 推理能效

推理能效是指推理卡在推理过程中，单位时间内消耗单位功耗，所完成的计算量，单位为万亿次浮点运算次数每瓦（TFLOPs/W）。推理能效的测试方法如表11所示。

表11 推理能效测量方法

度量指标	说明	测量方法
推理能效	推理卡单位时间内消耗单位功耗所完成的计算量。	计算公式为： $E = \frac{\text{size of (数据量)}}{T \times P}$ 其中： size of (*) ：在整个推理测试过程中，累计返回结果的任务数据量，计算并转化为浮点运算次数； T ：推理测试过程的用时； P ：推理测试过程的平均功率，可参照6.3.5.3的方法计算。

6.4 安全性要求

人工智能加速卡的安全性要求如下：

- 宜具备安全启动能力，逐级校验固件的完整性，确保设备自身安全性；
- 应具备基于硬件的安全加解密能力，加解密功能应符合 GM/T 0008-2012 的相关规定，芯片密码模块应符合 GB/T 37092-2018 及 GM/T 0028-2014 的相关规定；
- 可基于硬件的密码运算功能支持模型和数据加解密的保护能力；
- 宜具备 TRNG 真随机数发生器。

7 测试环境

7.1 测试对象

本文件将内置了人工智能加速卡的人工智能服务器作为被测对象，并在人工智能服务器上安装测试所需的操作系统、深度学习框架、加速卡驱动、软件栈、辅助测试工具等，用于测试人工智能加速卡所具备的功能，所能达到的性能，以及兼容性、可靠性、安全性等技术指标，并通过加载多种训练或推理任务负载，测试人工智能训练加速卡的训练性能和推理加速卡的推理性能。

7.2 测试组网

7.2.1 组网策略

测试系统包含被测服务器和测试机，测试机可选用桌面个人电脑，人工智能加速卡安装于被测服务器，通过在测试机上控制被测服务器，完成各项指标的测试。为了全面测试人工智能加速卡的技术指标，构建包含多台服务器节点的人工智能集群，每台服务器内安装1张以上人工智能加速卡。测试组网如图1所示。基于图1所示的组网环境，可分别开展单机单卡环境测试、单机多卡环境测试、集群环境测试。

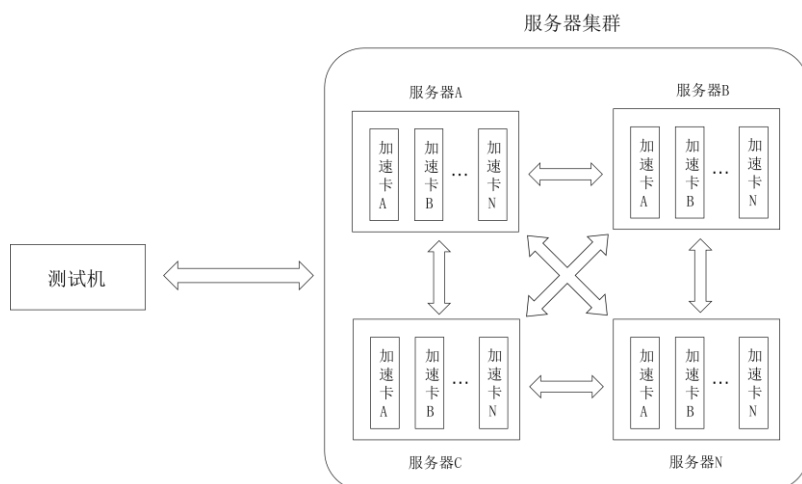


图1 人工智能服务器测试组网图

注 1：依据被测设备声明支持的节点内卡间互联方式，进行节点内加速卡的通信连接。

注 2：依据被测设备声明支持的节点间的网络通信方式，进行节点间的通信连接（如通过 RoCE、InfiniBand 等方式）。

7.2.2 单机单卡测试系统

通过相关设置，只启动服务器集群中的单台服务器中的单张人工智能加速卡，测试加速卡的相关技术指标。

7.2.3 单机多卡测试组网

通过相关设置，只启动服务器集群中的单台服务器中的多张人工智能加速卡，测试加速卡的相关技术指标。

7.2.4 集群测试组网

通过相关设置，启动服务器集群中的多个节点，以及服务器中的多张人工智能加速卡，测试加速卡的相关技术指标。

7.3 系统配置

7.3.1 软硬件配置

宜选用通用的x86或ARM架构服务器，服务器内安装需要进行测试的人工智能加速卡，并安装人工智能加速卡正常运行所需的驱动、软件栈等工具。

7.3.2 操作系统

被测设备应安装操作系统的开源或商业版本，常用的操作系统可参考附录A的表A.1所示。

7.3.3 深度学习框架

测试人工智能加速卡的各项技术指标时，需在被测设备中安装深度学习框架，以加载和运行各种人工智能算法模型，常用的深度学习框架可参考附录A的表A.2所示。

7.3.4 测试场景

本节给出测试人工智能训练加速卡和推理加速卡技术指标时宜加载的测试场景和任务，如表12所示。各类任务下用于测试加速卡技术指标的算法模型、数据集、模型的目标准确率等用例信息可参考附录A的A.3所示。

表12 测试场景

测试对象	模型类别	场景	任务
训练卡	传统模型	视觉类	图像分类
			目标检测
			图像分割
			视频分类
			文本识别
		语音类	语音识别
			声纹识别
	语音合成		
	推荐类	智能推荐	
	预训练模型	视觉类	
语言类			
推理卡	传统模型	视觉类	图像分类
			目标检测
			图像分割
			视频分类
			文本识别
		语音类	声纹识别
			语音合成
	推荐类	智能推荐	
	预训练模型	视觉类	
		语言类	

7.3.5 辅助工具

测试人工智能加速卡的各项技术指标，需要加速卡监控工具、功耗测试工具、通信带宽测试工具、加压测试工具等设备的配合，所需的辅助工具及用途如表13所示。

表13 辅助工具

序号	名称	用途
1	加速卡监控工具	加速卡各类指标的监控和读取
2	功耗测试工具	功耗测试
3	通信带宽测试工具	带宽测试
4	加压测试工具	稳定性测试

7.4 环境条件

除另外规定外，测试均在GB/T 9813.3-2017中规定的正常大气条件下进行，如表14所示。

表14 环境条件

序号	环境条件	数值范围
1	温度	15℃-35℃
2	相对湿度	25%-75%
3	大气压	86 kPa-106 kPa

8 测试方法

8.1 预置条件

测试人工智能加速卡的技术要求应满足如下预置条件：

- a) 被测设备电源供电正常；
- b) 被测设备安装稳定的商用 BIOS 版本及 BMC 版本；
- c) 被测设备预装业务操作系统并且各驱动程序安装正常；
- d) 被测设备预装容器化服务必要的组件，如容器化引擎、编排管理器、镜像仓库等；
- e) 被测设备安装加速卡驱动软件、计算引擎、开发环境等软件，且运行正常；
- f) 被测设备完成组网，可与测试机正常通信。

8.2 通用技术要求测试

8.2.1 基础配置管理功能测试

加速卡基础配置管理功能的测试方法如下：

- a) 通过测试机登录被测设备业务操作系统中；
- b) 使用命令查询加速卡的生产厂家、产品型号、序列号、固件版本、显存信息等基础配置信息；
- c) 验证相关信息是否可正常读取，并核对与加速卡实际配置及规格信息是否一致。

8.2.2 资产管理功能测试

加速卡资产管理功能的测试方法如下：

- a) 通过测试机登录被测设备的管理地址；
- b) 进入管理界面或带外命令，查看并读取加速卡的序列号信息；
- c) 登录到被测设备业务操作系统中，查看相应加速卡的序列号；
- d) 对比 b) 和 c) 中查询出来的加速卡序列号是否一致。

8.2.3 固件版本管理功能测试

加速卡固件版本管理功能的测试方法如下：

- a) 通过测试机登录被测设备的管理地址；
- b) 进入管理界面或带外命令，查看并读取加速卡的固件版本信息；
- c) 登录到被测设备业务操作系统中，查看相应加速卡的固件版本；
- d) 对比 b) 和 c) 中查询出来的加速卡固件版本信息是否一致。

8.2.4 功耗监控功能测试

加速卡功耗监控功能的测试方法如下：

- a) 通过测试机登录被测设备的管理地址；
- b) 进入管理界面或带外命令，查看并读取加速卡的功耗信息；
- c) 登录到被测设备业务操作系统中，查看相应加速卡的功耗值；
- d) 对比 b) 和 c) 中查询出来的加速卡功耗值差距是否在 5% 以内。

8.2.5 温度监控功能测试

加速卡温度监控功能的测试方法如下：

- a) 通过测试机登录被测设备的管理地址；
- b) 进入管理界面或带外命令，查看并读取加速卡当前温度信息；
- c) 登录到被测设备业务操作系统中，查看相应加速卡的当前温度值；
- d) 对比 b) 和 c) 中查询出来的加速卡温度值差距是否在 5% 以内。

8.2.6 加速卡操作系统适配测试

加速卡操作系统适配的测试方法如下：

- a) 通过测试机登录被测设备业务操作系统中；
- b) 执行操作系统检测脚本，检查业务操作系统中的版本和型号信息；
- c) 在此业务操作系统环境中，完成与加速卡所有技术要求对应的全部测试内容；
- d) 验证 b) 中获取的业务操作系统数据是否正确；
- e) 验证 c) 是否能够正常执行测试过程。

8.2.7 加速卡信创操作系统适配测试

加速卡信创操作系统适配的测试方法如下：

- a) 通过测试机登录被测设备业务操作系统中；
- b) 执行操作系统检测脚本，检查业务操作系统中的版本和型号信息，是否满足信创操作系统要求；
- c) 在此业务操作系统环境中，完成与加速卡所有技术要求对应的全部测试内容；
- d) 验证 b) 中获取的业务操作系统数据是否正确；
- e) 验证 c) 是否能够正常执行测试过程。

注：若 8.2.6 中使用的操作系统是信创操作系统，可复用测试结果。

8.2.8 与 CPU 通信能力的测试

加速卡与 CPU 之间通信能力的测试方法如下：

- a) 通过测试机登录被测设备业务操作系统中；
- b) 安装 CPU 与加速卡之间通信带宽的测试工具；
- c) 执行相关命令，通过 b) 中的测试工具获取 CPU 到加速卡之间的通信带宽，并记录所有带宽值；
- d) 执行相关命令，通过 b) 中的测试工具获取加速卡到 CPU 之间的通信带宽，并记录所有带宽值；
- e) 对 c) 和 d) 中获取的所有带宽值求平均值，得出 CPU 与加速卡之间的通信带宽性能 (GB/s)。

8.2.9 加速卡性能分析工具测试

加速卡支持性能分析工具的测试方法如下：

- a) 通过测试机登录被测设备业务操作系统中；

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/375310231022011231>