ICS 35.240.40
CCS A 11

团　体　标　准

T/BFIA 035—2024

金融业人工智能服务器应用技术要求

Technical requirements for artificial intelligence server in the financial industry

2024－11－29 发布　　2024－11－29 实施

北京金融科技产业联盟　　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》的规定起草。

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专利的责任。

本文件由北京金融科技产业联盟归口。

本文件起草单位：中国金融电子化集团有限公司、北京金安信息技术有限责任公司、中国建设银行股份有限公司、中国工商银行股份有限公司、中国民生银行股份有限公司、中国农业银行股份有限公司、中信银行股份有限公司、中国银行股份有限公司、上海浦东发展银行股份有限公司、平安银行股份有限公司、国家开发银行、招商银行股份有限公司、华为技术有限公司、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、中国长城科技集团股份有限公司、曙光信息产业股份有限公司、新华三技术有限公司、上海兆芯集成电路股份有限公司、飞腾信息技术有限公司、深圳市江波龙电子股份有限公司、四川华鲲振宇智能科技有限责任公司、第四范式（北京）技术有限公司。

本文件主要起草人：姜云兵、班廷伦、马国照、韩竺吾、常璐、刘东东、裴凯洋、毕伟光、龚郅凡、刁翔宇、林晨、朱昊志、甘政兵、高金鹏、蔡佳、宋辰、夏梦婷、孙朝斌、刘雪涛、钱学成、吴酋珉、白阳、王君、邸贺亮、颜培源、胡世珺、高云超、杨帆、薛石磊、张毅、杨景瑞、曹洵峰、刘东、刘胜龙、广文博、王桐桐、詹谦、顾伟。

引　言

在信息技术的全面应用创新趋势下，确保金融基础设施的安全稳定，是金融机构行稳致远关键之一。金融行业应用人工智能技术，是从数字化走向智能化的核心力量，更是金融机构智慧再造的关键载体。人工智能技术可自动化金融业务流程，提高效率和准确性。人工智能可帮助金融机构更好地识别和管理风险，并确保合规性。人工智能可通过分析大量数据，提供个性化的金融产品和服务。“算力、数据、算法、开放平台”是人工智能技术的核心内容，其中算力包括：人工智能芯片、人工智能设备等产品，提供金融机构使用高性能、低成本、绿色的人工智能算力是应用的关键目标。

金融业人工智能服务器种类繁多，主要有人工智能训练服务器、人工智能推理服务器、人工智能边缘服务器、人工智能服务器集群等，通过梳理金融业人工智能设备应用要求，有助于金融机构依据自身业务特点，针对性选购人工智能服务器，并采取适当、合理的管理措施和安全防护措施。为了加快金融服务智慧应用，特制定本文件。

金融业人工智能服务器应用技术要求

1 范围

本文件规定了人工智能服务器产品总体要求以及关键组件、兼容性、可靠性、迁移能力、供应链安全的要求，给出了性能测试方案和安全可控特性的说明。

本文件适用于人工智能服务器产品的设计、开发、生产、服务保障等环节，该文件可作为各金融行业等相关单位进行信息系统人工智能改造升级时参考。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

服务器 server

信息系统的重要组成部分，是信息系统中为客户端计算机提供特定应用服务的计算机系统，由硬件系统（处理器、存储设备、网络连接设备等）和软件系统（操作系统、数据库管理系统、应用系统）组成。

[来源：GB/T 9813.3—2017，3.1]

3.2

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。

注：本文件中，在不引起误解的语境中，将人工智能加速卡简称为加速卡。

[来源：GB/T 42018—2022，3.6]

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计算加速能力的服务器，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。

注3：本文件中，在不引起误解的语境中，将人工智能服务器简称为服务器。

[来源：GB/T 42018—2022，3.5]

3.4

人工智能加速处理器 artificial intelligence accelerating processor

具备适配人工智能算法的运算微架构，能够完成人工智能应用加速运算处理的集成电路元件。

注：本文件中，在不引起误解的语境中，将人工智能加速处理器简称为人工智能处理器。

[来源：GB/T 42018—2022，3.8，有修改]

3.5

训练 training

利用训练数据，基于机器学习算法，建立或改进机器学习模型参数的过程。

[来源：ISO/IEC 22989:2021，3.2.15]

3.6

推理 inference

计算机根据已知信息进行分析、分类或诊断，做出假设，解决问题或者给出推断的过程。

注：人工智能领域的推理包括逻辑推理、机器学习推理等。

[来源：GB/T 42018—2022，3.12]

3.7

安全可控 controllability for security

信息技术产品具备的保证其应用方数据支配权、产品可控权、产品选择权等不受损害的属性。

[来源：GB/T 36630.1—2018，3.2]

3.8

深度学习 deep learning

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。

注：深度学习是机器学习的分支。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

BMC：基板管理控制器（Baseboard Management Controller）

CPU：中央处理器（Central Processing Unit）

DDR：双倍速率（Double Data Rate）

ECC：错误检测和纠正（Error Correcting Code）

FLOPS，每秒浮点运算次数（Floating-point Operations Per Second）

GE：千兆以太网（Gigabit Ethernet）

GPU：图形处理器（Graphics Processing Unit）

HBM：高带宽内存（High Bandwidth Memory）

HDD：硬盘有机械硬盘（Hard Disk Drive）

NPU：嵌入式神经网络处理器（Neural-network Processing Units）

NVMe：非易失性内存协议（Non Volatile Memory Express）

PCIE：一种高速串行计算机扩展总线标准（Peripheral Component Interconnect Express）

RAID：磁盘阵列（Redundant Arrays of Independent Disks）

RDIMM：带寄存器的双线内存模块（Registered Dual In-line Memory Module）
RoCE：基于融合以太网的 RDMA（RDMA over Converged Ethernet）
SSD：固态硬盘（Solid State Disk 或 Solid State Drive）
TOPS：每秒运算处理 Tera 次（Tera Operations Per Second）

5 金融 AI 服务器产品总体要求

5.1 概述

按照AI服务器的使用类型和部署方式分为AI训练服务器、AI推理服务器、AI机柜式服务器等。

5.2 AI 训练服务器

5.2.1 通用要求

AI 训练服务器满足以下要求：
a）应支持提供的人工智能计算加速方式如下：
——通过扩展设备，如人工智能加速卡等；
——通过加速或配套的扩展加速模组，如人工智能加速电路等；
——通过集成人工智能处理器的方式，如 SOC 等。
b）外部存储应支持 SATA SSD 或 NVMe SSD；
c）应支持连接以太网（包括 RoCE 网络）接口；
d）整机应至少配置 2 个人工智能处理器，每个人工智能处理器的显存应不低于 48GB；每个人工智能处理器的 FP16 算力应不小于 70 TFLOPS。

5.2.2 推荐性功能

AI 训练服务器推荐性功能如下：
a）CPU 每个核心 L1 I-cache 宜不小于 32KB，L1 D-cache 宜不小于 32KB，L2 cache 宜不小于 512KB；
b）L3 cache 容量宜不小于 48MB；
c）宜支持 DDR4 及以上版本的内存，DDR 通道宜不少于 8 个；
d）外部存储宜支持 SATA 硬盘设置 RAID 0/1/10/5/50/6/60；
e）宜支持连接以太网（如 RoCE 网络）接口或 Infiniband 网络等接口；
f）CPU 宜支持机密计算，支持国密技术，能够扩展支持 GPU/NPU 的机密计算方案；
g）宜支持 ECC 1bit 纠错，ECC 2bit 报错；
h）训练服务器浮点算力：整机宜至少配置 8 个人工智能处理器。每个人工智能处理器的 FP32 算力宜不小于 80 TFLOPS ，FP16 算力宜不小于 300 TFLOPS，BF16 算力宜不小于 300 TFLOPS，INT8 算力宜不小于 600 TOPS；每个人工智能处理器显存宜不低于 64GB；带宽不低于 1600GB/S；
i）整机服务器内人工智能处理器双向通信带宽宜不小于 380 GB/s；宜支持多台服务器间互联，不小于 8*200GE ROCE。
训练 AI 服务器性能测试方案见附录 A。

5.3 AI 推理服务器

5.3.1 通用要求

AI推理服务器的要求如下：

a）应配备三级缓存，容量不应低于 16MB；
b）应支持 DDR4 或 LPDDR4 及以上版本的内存；
c）应兼容 PCIe 4.0 及更低版本的 PCIe 协议；
d）应能连接并使用 25GE、10GE 或 GE 等接口；
e）中心推理服务器，所安装推理卡，每卡要求如下：
——INT8 算力应不小于 140 TOPS，FP16 应不小于 70 TFLOPS；
——显存应不低于 24GB。
f）边缘推理服务器，所安装推理卡，每卡要求如下：
——INT8 算力应不小于 100 TOPS，FP16 应不小于 50 TFLOPS；
——应支持不小于 80 路（1080P 30 帧/秒）视频解码（视频格式如 H.264/H.265）；
——应支持不小于 24 路（1080P 30 帧/秒）视频编码（视频格式如 H.264/H.265）。

5.3.2 推荐性功能

AI 推理服务器推荐性功能如下：
a）宜能通过自带固件，如通过 BMC 等，或其他外接部件监控系统参数；
b）宜支持 INT8 运算；
c）中心推理服务器所安装的推理卡，每卡规定如下：
——INT8 算力宜不小于 560 TOPS，FP16 算力宜不小于 280 TFLOPS；每卡显存宜不低于 64GB，带宽宜不低于 1600GB/S；
d）边缘推理服务器所安装的推理卡，每卡规定如下：
——INT8 算力宜不小于 140 TOPS，FP16 宜不小于 70TFLOPS；
——宜支持不小于 128 路（1080P 30 帧/秒）视频解码（视频格式如 H.264/H.265）；
——宜支持不小于 24 路（1080P 30 帧/秒视频编码（视频格式如 H.264/H.265）。
AI 推理服务器的性能测试方案见附录 B。

5.4 AI 机柜式服务器

5.4.1 通用要求

对AI机柜式服务器的要求如下：
a）应支持通过安全可靠测评的 CPU 处理器，宜集成了 DDR4 或 DDR5、PCIe3.0 以上、100GE、25GE、10GE、GE 等接口，提供完整的 SOC 功能；
——应支持不低于 48 核，单核主频率应 2.6GHz 以上；
——应兼容安全可控的 CPU 架构；
——应支持收集 CPU 状态；
b）应支持面向深度学习训练的高性能多核 GPU/NPU 等处理器；
c）单台计算节点应支持不小于 8 个 AI 处理器，能够最大限度地提高多线程应用的并发执行能力；
d）应最小支持不少于 24 条 DDR4/5 ECC 内存，内存支持 RDIMM，支持最小提供 1024GB 内存容量；
e）应支持多种灵活的硬盘配置方案，提供了弹性的、可扩展的存储容量空间，满足不同存储容量的需求和升级要求。

5.4.2 推荐性功能

单柜宜支持放置最多 8 个计算节点，单个计算节点宜提供不低于 3000TFLOPS 的 FP16 算力，或不低于 750TFLOPS 的 FP32 算力。

6 金融 AI 服务器关键组件要求

6.1 硬件要求

6.1.1 通用计算芯片

6.1.1.1 通用要求

对通用计算芯片的要求如下：

a）应支持安全可控的处理器；

b）应支持多个内存通道。

6.1.1.2 推荐性功能

宜支持芯片级的机密计算和安全加速等安全功能。

6.1.2 AI 加速芯片

AI 加速芯片通用要求如下：

应配置安全可控的 AI 加速芯片。AI 芯片、AI 加速卡、AI 服务器的安全可控特征见附录 C。

6.1.3 AI 加速卡

6.1.3.1 通用要求

对AI加速卡的要求如下：

a）应支持程序配置、使用、管理 AI 加速卡；

b）应支持加速卡扩展，接口类型应具备通用性。

6.1.3.2 推荐性功能

AI加速卡的推荐性功能如下：

a）单卡 INT8 算力峰值宜不小于 280TOPS；

b）单卡 FP16 算力峰值宜不小于 140TFLOPS 算力；

c）宜支持视频硬件编解码功能；

d）宜支持 AI 加速卡的管理功能，如设备型号识别、设备温度获取等。

6.1.4 供电单元

对供电单元的通用要求如下：

a）应支持双路及以上电源冗余输入；

b）应支持通信接口，具备故障报警，提供电源运行状态监控；

c）应支持安全可控供电单元；

d）应支持高压直流功能。

6.1.5 散热单元

6.1.5.1 通用要求

对散热单元的要求如下：

a）应支持系统冗余散热，如通过风冷、液冷等散热方式实现系统散热；

b）应支持风扇调速，根据调速策略自动调整风扇转速。

6.1.5.2 推荐性功能

宜支持模块化设计，支持风扇模块热更换。

6.2 软件要求

6.2.1 操作系统

6.2.1.1 通用要求

应具备来自版权所有方合法的最终用户使用授权且功能正常。

6.2.1.2 推荐性功能

宜支持安全可控的操作系统。

6.2.2 AI 基础软件

6.2.2.1 AI 芯片使能软件[1)]

对 AI 芯片使能软件的通用要求以下：

a）AI 芯片使能软件应具备人工智能软件加速库（算子）的集合，提供对于深度学习的计算优化功能；

b）应通过提供多层次的编程接口，支持用户快速构建 AI 应用和业务；

c）应提供基于 C/C++等语言的算子开发接口，使用户具有自定义算子开发的能力；

d）AI 产品应使用 k8s 等进行算力资源的运维管理，并提供 AI 产品主要指标的监测能力。

6.2.2.2 AI 开发框架

6.2.2.2.1 通用要求

对AI开发框架的要求如下：

a）提供 AI 软件适配能力，应支持国内外深度学习框架；

b）应至少支持 1 种深度学习或分布式框架，包括但不限于 MindSpore、TensorFlow、PyTorch、PaddlePaddle 等。

6.2.2.2.2 推荐性功能

AI开发框架的推荐性功能如下：

a）宜支持常见的视觉分析、NLP 和语音识别功能；

b）视觉分析宜支持 resnet50、yoloV5 等神经网络；

c）NLP 宜支持 bert、Transformer 等神经网络；

d）语音识别宜支持 tacotron2、waveRNN、flyspeech 等神经网络。

7 金融 AI 服务器兼容性要求

7.1 通用要求

1) 使能软件是让芯片具有特定功能。

对 AI 服务器系统兼容性的要求如下：

a）服务器的兼容性主要包括部件兼容性和操作系统兼容性。部件兼容性应保障服务器中各部件的有效运行，包括 CPU、内存、网卡、HDD 硬盘、SSD 盘/卡等；操作系统兼容性主要保障服务器硬件与操作系统的有效配合；

b）固件升级前后应兼容官方版本。

7.2 推荐性功能

AI服务器系统兼容性的推荐性功能如下：

a）宜支持通用的 AI 开发框架；

b）宜支持 AI 芯片使能软件；

c）宜支持模型转换、模型可解释、功能调试、性能调优等。

8 金融 AI 服务器可靠性要求

8.1 通用要求

对整机及部件可靠性的要求如下：

a）应支持服务器关键部件冗余设计，包括 HDD/SSD、电源、风扇等部件；

b）应支持服务器关键部件异常报警功能，包括 CPU、HDD/SSD、内存等部件；

c）应支持服务器关键部件故障定位机制，包括 CPU、HDD/SSD、内存等部件；

d）应支持服务器 HDD/SSD、电源、风扇等部件热插拔功能；

e）应支持内存可靠性技术，如内存查错、纠错等；

f）应支持 CPU、内存条、GPU、网卡等的现场扩容替换以及故障件替换。

8.2 推荐性功能

对整机及部件可靠性的推荐性功能如下：

a）宜支持内存可靠性技术，如内存冗余备份等；

b）宜支持服务器关键部件故障隔离机制，包括 CPU、HDD/SSD、内存等部件；

c）宜支持 I/O 模块在线隔离或更换；

d）宜支持 CPU、内存条、GPU、网卡等的现场扩容替换以及故障件替换；

e）宜支持部件的故障自动修复功能。

9 算子模型迁移能力要求

9.1 算子迁移开发能力要求

9.1.1 通用要求

应提供三方算子迁移工具，支持异常自动检测，支持内存检测和线程检测。

9.1.2 推荐性功能

对算子迁移开发能力的推荐性功能如下：

a）宜提供完备的功能调试功能，支持孪生调试和上线调试；

b）宜提供极限性能分析评估能力，宜支持核函数和热点函数等性能分析；

c）宜支持容错能力，如断点续训能力。

9.2 加速库兼容要求

9.2.1 通用要求

对加速库兼容的要求如下：

a）应支持三方社区及加速卡兼容，支持机器视觉领域算子库和套件、图像分类、图像检测套件、语义分割等；

b）应支持分布式并行加速库（deepspeed、MegatronLM、Triton 等）；

c）应支持多领域开放套件，包括但不限于自然语言处理类模型套件（hugging face）；机器视觉类套件（OpenMMLab 等）；

d）AI 算力处理卡应支持图像、视频、文字、语音等多种数据分析与推理计算，支持金融场景；

e）应支持离线推理模型对接，如 ONNX。

9.2.2 推荐性功能

加速库兼容的推荐性功能如下：

a）宜支持增量/迁移学习；

b）宜支持跨节点预训练等。

9.3 大模型迁移能力要求

9.3.1 通用要求

对大模型迁移能力的要求如下：

a）应支持分布式环境下模型的保存和重加载，支持多维混合并行；

b）应提供脚本迁移能力，提供分钟级完成度评估能力；

c）应支持行代码级转换能力，提供脚本转换；

d）应提供低代码自动并行，支持以单卡开发视角实现千亿参数模型自动并行；

e）应提供大模型加速库，支持面向 transformer 结构的软硬件协同优化；

f）应提供模型压缩组件，支持面向大模型的自动压缩加速。

9.3.2 推荐性功能

对大模型迁移能力的推荐性功能如下：

a）宜提供在确保安全隐私的前提下，支持大模型跨域协同训练能力，如跨区域、跨资源池；

b）宜提供模型微调流程模板，集成低参微调算法，支持部分参数微调，降低存储资源占用，提升微调效率；

c）宜支持模型精度调优，提供数据 Dump[2)]，溢出检测，精度比对能力；

d）宜提供数据采集解析可视化对比工具，支持计算性能调优，计算通信比和算子耗时分析；

e）宜提供基于时间线和统计图的可视化分析，提升计算瓶颈识别定位效率；

f）宜提供面向典型大模型结构负载的性能 Benchmark。

10 金融 AI 服务器供应链安全要求

2） Dump 是转储，代指将某种数据以某种格式进行转储或导出过程。

10.1 针对产品要求

对AI服务器供应链中的产品通用要求如下：

a）设计、开发、生产等关键环节，应实施必要的安全防护措施；

b）应不存在未声明功能和已知的安全风险；

c）应具备供应链安全性和持续稳定性说明。

10.2 针对 AI 服务器供应方的要求

对AI服务器供应链的供应方通用要求如下：

a）AI 服务器供应方应具备产品研发设计、生产制造、供应保障、售后维护相匹配的人员和工作环境；

b）AI 服务器供应方应具备产品定制开发能力，能够基于自身产品构建产业生态，保持生态开放性、透明性、满足各种应用场景需求；

c）AI 服务器供应方应具备漏洞响应等能力和管理机制；

d）AI 服务器供应方应具备及时有效的售后服务能力与管理机制；

e）AI 服务器供应方应满足安全可控要求；

f）服务器供应方应在官方网站提供服务器部件兼容性查询功能；

g）服务器供应方应在官方网站提供服务器操作系统兼容性列表；

h）AI 服务器供应方不应无故停止已约定产品供应或停止已约定服务。

附 录 A
（资料性）
训练 AI 服务器性能测试方案

A.1 图像检测场景

对图像检测场景，基于PyTorch框架YoloV5m-6.0模型的测试方法见表A.1。

表A.1 PyTorch框架下YoloV5m-6.0模型训练性能测试

测试环节	测试内容说明
测试目的	测试被测设备在 PyTorch 框架下，基于 coco2017 数据集使用 YoloV5m-6.0 模型进行训练的性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备为稳定的商用 BIOS 版本及 BMC 版本 4.设备已安装规定操作系统并且驱动程序安装正常 5.业务系统中已安装加速卡对应软件套件、Driver 6.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 7.业务系统中已导入 coco2017 数据集 8.业务系统中已导入训练代码， 9.业务系统中已导入测试容器镜像
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 说明：所有参数均不允许修改。 3.记录日志获取训练时间及每秒处理的样本数（samples/s），日志打印的模型准确率要求达到训练目标准确率
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤 3 中，从日志中获取每秒处理的样本数（samples/s），模型准确率能够达到目标准确率。
测试说明	—
测试结果	以实际测试为准
测试说明	1.目标准确率 map@0.5:63% 2.性能测试方法：从第 2 个 epoch 开始进行性能统计 （训练稳定），取 10 个 epoch 的平均吞吐性能。

A.2 图像分类场景

对图像分类场景，基于pytorch框架Resnet50模型的测试方法见表A.2。

表A.2 PyTorch框架下Resnet50模型训练性能测试

测试环节	测试内容说明
测试目的	测试被测设备在 PyTorch 框架下，基于 CIFAR100 数据集使用 Resnet50 模型进行训练的性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器

表A.2 PyTorch框架下Resnet50模型训练性能测试（续）

测试条件	3.设备为稳定的商用 BIOS 版本及 BMC 版本 4.设备已安装规定操作系统并且驱动程序安装正常 5.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 6.业务系统中已导入 CIFAR100 数据集 7.业务系统中已导入训练代码 7.业务系统中已导入测试容器镜像
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 3.记录日志获取训练时间及每秒处理的样本数（samples/s），日志打印的模型准确率要求达到训练目标准确率
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤 3 中，从日志中获取每秒处理的样本数（samples/s），模型准确率能够达到目标准确率
测试说明	1.目标准确率 acc@1:79.90%
测试结果	以实际测试为准

A.3 OCR场景

对OCR场景，基于MindSpore框架下DBNet模型的测试方法见表A.3。

表A.3 MindSpore框架下DBNet模型训练性能测试

测试环节	测试内容说明
测试目的	测试被测设备在 MindSpore 框架下，基于 ICDAR2015 数据集使用 DBNet 模型进行训练的性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备为稳定的商用 BIOS 版本及 BMC 版本 4.设备已安装规定操作系统并且驱动程序安装正常 5.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 6.业务系统中已导入 ICDAR2015 数据集 7.业务系统中已导入训练代码 8.业务系统中已导入测试容器镜像
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 3.记录日志获取训练时间及每秒处理的样本数（samples/s），日志打印的模型准确率要求达到训练目标准确率
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤 3 中，从日志中获取每秒处理的样本数（samples/s），模型准确率能够达到目标准确率
测试说明	1.目标准确率 Hmean:70.00%
测试结果	以实际测试为准

A.4 NLP场景

对NLP场景，基于PyTorch框架下BertBase模型的测试方法见表A.4。

表A.4 PyTorch框架下BertBase模型训练性能测试

测试环节	测试内容说明
测试目的	测试被测设备在 PyTorch 框架下，基于人民日报数据集，使用 BertBase 模型进行训练的性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备为稳定的商用 BIOS 版本及 BMC 版本 4.设备已安装规定操作系统并且驱动程序安装正常 5.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 6.业务系统中已导入人民日报数据集 7.业务系统中已导入代码 8.业务系统中已导入预训练模型 9.业务系统中已导入测试容器镜像
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 3.记录日志获取训练时间及每秒处理的样本数（samples/s），日志打印的模型准确率要求达到训练目标准确率
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤 3 中，从日志中获取每秒处理的样本数（samples/s），模型准确率能够达到目标准确率
测试说明	1.目标准确率 micro-F1:89.00%
测试结果	以实际测试为准

A.5 语音场景

对语音场景，基于Pytorch框架下espnet-conformer模型的测试方法见表A.5。

表A.5 PyTorch框架下espnet-conformer模型训练性能测试

测试环节	测试内容说明
测试目的	测试被测设备在 PyTorch 框架下，基于 aishell-1 数据集使用 espnet-conformer 模型进行训练的性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备为稳定的商用 BIOS 版本及 BMC 版本 4.设备已安装规定操作系统并且驱动程序安装正常 5.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 6.业务系统中已导入 aishell-1 数据集 7.业务系统中已导入训练代码 7.业务系统中已导入测试容器镜像
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令采用混合精度进行模型训练单机 8 卡训练 3.记录日志获取训练时间及每秒处理的样本数（samples/s），日志打印的模型准确率要求达到

表A.5 PyTorch框架下espnet-conformer模型训练性能测试（续）

测试过程	训练目标准确率
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤 3 中，从日志中获取性能数据：每秒处理的语音数量，模型准确率能够达到目标准确率。
测试记录	—
测试结果	以实际测试为准
测试说明	1. 目标准确率：est Corr=95.4 2. 性能测试方法：数据集数量 * 3（3 种倍速） * epoch / 训练总时间。

A.6 大模型训练场景

对大模型训练场景，基于Pytorch框架下ChatGLM-6B模型的测试方法见表A.6.1。

表A.6.1 Pytorch框架下ChatGLM-6B模型训练性能测试

测试环节	测试内容说明	
测试目的	测试被测设备在 PyTorch 框架下，基于 ADGEN 数据集，使用 ChatGLM-6B 模型在单机 8 卡上进行训练的性能	
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备已安装规定操作系统并且驱动程序安装正常 4.业务系统中已安装 PyTorch 框架、Python 依赖和加速卡对应软件套件 5.业务系统中已导入 ADGEN 数据集 6.业务系统中已导入代码 7.业务系统中已导入预训练模型	
模型配置	配置项	NPU 参数
	bos_token_id	150004
	eos_token_id	150005
	hidden_size	4096
	inner_hidden_size	16384
	Layer norm_epsilon	1e-05
	num_attention_heads	32
	num_layers	28
	max_sequence_length	2048
	Vocab_size	150528
	torch_dtype	float16
	position_encoding_2d	true
	use_cache	true
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 3.记录日志，解析每秒处理的样本数（tokens/s），loss 下降趋势	
预期结果	1.步骤 1 中，所有人工智能加速卡均在位且运行正常，无告警；	

表A.6.1 Pytorch框架下ChatGLM-6B模型训练性能测试（续）

预期结果	2.步骤3中，从日志中获取每秒处理的样本数（tokens/s），loss曲线随迭代步数下降
测试说明	—
测试结果	以实际测试为准

对大模型训练场景，基于Pytorch框架下LLaMA-13B模型的测试方法见表A.6.2。

表A.6.2 Pytorch框架下LLaMA-13B模型训练性能测试

测试环节	测试内容说明	
测试目的	测试被测设备在PyTorch框架下,基于alpaca-data-conversation.json数据集,使用LLaMA-13B模型在两机16卡上进行训练的性能	
测试条件	1.设备电源供电正常 2.使用本地管理PC连接到服务器 3.设备已安装规定操作系统并且驱动程序安装正常 4.业务系统中已安装PyTorch框架、Python依赖和加速卡对应软件套件 5.业务系统中已导入alpaca-data-conversation.json数据集 6.业务系统中已导入代码 7.业务系统中已导入预训练模型	
模型配置	配置	NPU参数
	Per_device_bz	1
	Gradient_accumulation_steps	16
	Save_steps	120000
	Save_total_limit	1
	Learning rate	2.00E-05
	Weight decay	0
	Warmup ratio	0.03
	Lr scheduler type	Cosine
	Fp16	TRUE
	Model_max_length	1024
	Gradient checkpointing	TRUE
	Lazy preprocess	TRUE
	Zero	2
	FSDP	FALSE
测试过程	1.远程通过命令行登录被测试服务器，执行命令查询当前人工智能加速卡的状态及健康状况 2.执行命令进行模型训练 3.记录日志，解析每秒处理的样本数（tokens/s），loss下降趋势	
预期结果	1.步骤1中，所有人工智能加速卡均在位且运行正常，无告警； 2.步骤3中，从日志中获取每秒处理的样本数（tokens/s），loss曲线随迭代步数下降	
测试说明	—	
测试结果	以实际测试为准	

附 录 B
（资料性）
推理 AI 服务器性能测试方案

B.1 视频编解码测试

B.1.1 视频解码性能测试

对视频解码场景，推理AI服务器性能测试方法见表B.1.1。

表B.1.1 视频解码性能测试

测试环节	测试内容说明
测试目的	通过一段视频测试芯片视频硬解码性能
测试条件	1.设备电源供电正常； 2.使用本地管理 PC 连接到服务器； 3.设备已安装 Linux 操作系统并且驱动程序安装正常； 4.测试软件、测试工具已安装完成； 5.准备测试视频 video_h264_1080p.mp4、video_h265_1080p.mp4； 6.视频要求：输入视频的高度，默认值为 1080。
测试过程	远程通过命令行登录被测试服务器； 执行命令进行测试； 记录视频硬解码吞吐率。 监控 CPU 使用率，确保 CPU 不参与解码过程。
预期结果	1.记录 H.264 性能值，包括 fps 帧率信息和最大支持路数； 2.记录 H.265 性能值，包括 fps 帧率信息和最大支持路数。
测试记录	—
测试结果	以实际测试为准
测试说明	1.解码软件不限制； 2.#默认视频解码配置为 N（不限制）路视频并行执行，可以通过 ffmpegN（不限制）实时显示的稳定（上下浮动不超过 1%）的 fps 帧率信息乘以 N 路得到最终累加的总帧率，即为该格式视频的解码性能值； 3.#提供的解码测试视频均为 1920x1080； #H.264 解码性能测试视频； #H.265 解码性能测试视频

B.1.2 视频编码性能测试

对视频编码场景，推理AI服务器性能测试方法见表B.1.2。

表B.1.2 视频编码性能测试

测试环节	测试内容说明
测试目的	通过一段视频测试芯片视频编码性能
测试条件	1.设备电源供电正常； 2.使用本地管理 PC 连接到服务器

表B.1.2 视频编码性能测试（续）

测试条件	3.设备已安装 Linux 操作系统并且驱动程序安装正常； 1.测试软件、测试工具已安装完成； 2.准备测试视频 input_1920x1080_50.yuv。 3.视频要求：输入视频的高度，默认值为 1080。 下列参数要求统一（参考 ffmpeg 命令参数：-s:v 1920x1080 -rc_mode VBR -r 25 -g 250 -b:v 8m）
测试过程	远程通过命令行登录被测试服务器； 执行命令进行测试； 记录视频硬解码吞吐率。 监控 CPU 使用率，确保 CPU 不参与编码过程。
预期结果	记录 H.264 性能值，包括 fps 帧率信息和最大支持路数。 记录 H.265 性能值，包括 fps 帧率信息和最大支持路数。
测试记录	—
测试结果	以实际测试为准
测试说明	1.编码路数和解码软件不限制； 2.默认视频编码配置为 N（不限制）路 1080P 并行执行，可以通过 ffmpeg（不限制）等工具显示的稳定（上下浮动不超过 1%）的 fps 帧率信息乘以 N 路得到最终累加的总帧率，即为 H.264/H.265 的编码性能值。 3.其中编码是基于 yuv420P 或 yuv420 下的 1080p 视频：nput_1920x1080_50.yuv 或者基于 yuv 转为其他格式

B.2 推理应用性能测试

B.2.1 图像检测场景

对图片检测场景，基于YoloV5s模型推理精度，推理AI服务器性能测试方法见表B.2.1。

表B.2.1 图像检测推理性能测试

测试环节	测试内容说明
测试目的	测试 YOLOv5s 模型推理性能
测试条件	1.设备电源供电正常 2.使用本地管理 PC 连接到服务器 3.设备为稳定的商用 BIOS 版本 4.设备已安装 Linux 操作系统并且驱动程序安装正常 5.被测服务器安装该芯片的相关库与驱动版本，并记录版本信息 6.使用 COCO2017 val 数据集 7.下载开源链接，并做模型量化与转换。
测试过程	1.远程通过命令行登录被测试服务器 2.使用 benchmark 工具进行部署 3.执行下列命令进行图片推理：./yolov5s_benchmark.sh（厂商提供） 4.说明：所有参数定下来后不允许修改 5.记录日志并获取平均每秒处理图片的性能数据（图片处理速度 images/sec）、图片数量、推理时间等。

表 B.2.1 图像检测推理性能测试（续）

预期结果	1.正常推理，从日志中获取并记录平均每秒处理图片的性能数据； 2.map@0.5>54。
测试记录	—
测试结果	以实际测试为准
测试说明	—

B.2.2 图像分类场景

对图片分类场景，基于YResNet50模型推理精度，推理AI服务器性能测试方法见表B.2.2。

表B.2.2 图片分类推理性能测试

测试环节	测试内容说明
测试目的	测试 ResNet50 模型推理性能
测试条件	1.设备电源供电正常； 2.使用本地管理 PC 连接到服务器； 3.设备为稳定的商用 BIOS 版本； 4.设备已安装 Linux 操作系统并且驱动程序安装正常； 5.被测服务器安装该芯片的相关库与驱动版本，并记录版本信息； 6.使用 imagenet2012 val 数据集； 7.使用 Pytorch torchvision 中提供的 ResNet50 的模型，模型下载链接为下载开源链接，做模型量化与转换；
测试过程	1.远程通过命令行登录被测试服务器； 2.使用 benchmark 工具进行部署； 3.执行下列命令进行图片推理：./resnet50_benchmark.sh（厂商提供）； 4.记录日志并获取平均每秒处理图片的性能数据（图片处理速度 images/sec）、图片数量、推理时间等。
预期结果	1.正常推理，从日志中获取并记录平均每秒处理图片的性能数据； 2.记录准确率，要求 TOP1 准确率大于 74.5，TOP5 准确率大于 91。
测试记录	—
测试结果	以实际测试为准
测试说明	—

B.2.3 自然语言处理场景

对自然语言处理场景，基于BertBase模型推理精度，推理AI服务器性能测试方法见表B.2.3。

表B.2.3 自然语言处理推理性能测试

测试环节	测试内容说明
测试目的	测试 Bert-base 模型推理性能
测试条件	1.设备电源供电正常； 2.使用本地管理 PC 连接到服务器； 3.设备为稳定的商用 BIOS 版本； 4.设备已安装 Linux 操作系统并且驱动程序安装正常； 5.被测服务器安装该芯片的相关库与驱动版本，并记录版本信息；

表B.2.3 自然语言处理推理性能测试（续）

测试条件	6.使用数据集进行测评；
测试条件	7.使用模型下载链接为下载开源链接，并做模型量化与转换； 8.使用容器部署方式； 9.统一把数据集都先下载准备好，统一为离线推理。
测试过程	1.远程通过命令行登录被测试服务器； 2.使用 benchmark 工具进行部署； 3.执行下列命令进行样本推理：./bert_benchmark.sh（厂商提供）； 4.记录日志并获取平均每秒处理样本的性能数据（样本处理速度 samples/sec）、样本数量、推理时间等。
预期结果	1.正常推理，并输出模型吞吐率数据，记录精度数据； 2.评价指标：F1 或者 Precision，或 ACC>89.5。
测试记录	—
测试结果	以实际测试为准
测试说明	—

B.2.4 大模型场景

对大模型场景，基于Stable Diffusion模型，推理AI服务器性能测试方法见表B.2.4。

表B.2.4 大模型推理性能测试

测试环节	测试内容说明
测试目的	测试单卡 Stable Diffusion 2.1 模型推理性能
测试条件	1.设备电源供电正常； 2.使用本地管理 PC 连接到服务器； 3.设备为稳定的商用 BIOS 版本； 4.设备已安装 Linux 操作系统并且驱动程序安装正常； 5.被测服务器安装该芯片的相关库与驱动版本，并记录版本信息； 6.数据集准备； 7.获取精度评估脚本； 8.git 获取评估的 clip 模型权重； 9.执行模型转换。
测试过程	1.设置 batchsize，推理精度使用 FP16/BF16（2 选 1），模型输出图片大小 512*512，50 次迭代；其他模型参数使用默认参数，运行测试脚本，计算模型推理吞吐率： 吞吐率=BatchSize*1/推理时延（单位 s）。 2.测试精度： （1）使用测试设备，Parti 数据集进行推理，得到生成的图片。 （2）按评估脚本所需的 image_info 进行输出： json 格式，list 对象。list 每个成员为 3 个 key 的字典：key’images’表示图片路径，key’category’表示输出的图片类别，key’prompt’表示提示文本） （3）image_info 保存成 json 文件。 （4）调用 clip_score.py（支持 CPU/GPU 执行），计算精度得分。
预期结果	1. 正常推理，并输出模型吞吐率数据，记录精度数据；

表 B.2.4 大模型推理性能测试（续）

预期结果	2.精度评价指标： ClipScore≥0.377。 3.如本模型无法正常运行或无法执行所有测试组合用例或精度未达到要求，则判定为不通过。

附 录 C
（资料性）
AI 服务器安全可控特性说明

C.1 AI芯片

C.1.1 核心技术说明

对AI芯片说明如下：

a）提供自行研发或有知识产权的架构设计或永久授权和具备自研能力，能举证架构设计文件；
b）能研制加速器内核，包含指令集、微架构。微架构和 RTL 代码开发，能举证；
c）能配套研制图像预处理、编解码硬件部件，并提供相关配套软件或 API；
d）能独立设计、研制加速库或算子实现库，开发由加速器提供商自行完成；
e）能提供核函数开发工具，训练加速库和推理加速引擎；
f）配备或搭配的 CPU、OS 满足安全可控的要求；
g）能研制或配套机器学习系统集成开发环境（包括但不限于编译构建平台、代码管控溯源、版本管理、漏洞管理、编码规范扫描等）。

C.1.2 供应链可持续说明

对人工智能加速器供应链可持续的说明如下：

a）使用有知识产权的 AI 处理器 IP 核；
b）加速库或算子实现库具备知识产权接口的设计，且版本演进。

C.2 AI加速卡

C.2.1 核心技术说明

AI 加速卡芯片符合 A.1.1 的 AI 芯片。

C.2.2 供应链可持续说明

对人工智能加速卡供应链可持续的说明如下：

a）AI 加速卡供应链可持续；
b）具备知识产权卡架构设计。

C.3 AI服务器核心技术说明

AI 服务器符合以下组成方面：

a）能独立设计 AI 服务器体系架构，能举证架构设计文件；
b）使用符合要求的 AI 芯片；
c）配备图像预处理，编解码硬件部件，该厂商提供相关配套软件或 API；
d）提供或者兼容至少 1 种集群高速互联部件和协议
e）兼容至少 2 种深度学习框架；
f）兼容至少 2 种操作系统；

g）提供或配套的应用程序开发 IDE 满足安全可控；

h）提供或配套的应用程序支持库满足安全可控要求，其功能包含：虚拟化调度，云、边、端协同，视觉场景 SDK 等。

参 考 文 献

[1] GB/T 36630.1—2018 信息安全技术 信息技术产品安全可控评价指标 第1部分：总则

[2] IEEE P 3142 Recommended Practice on Distributed Training and Inference for Large-scale Deep Learning Models

[3] IEEE Std 2937-2022 IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems

[4] ISO/IEC TR 24030 Information technology — Artificial intelligence (AI) — Use cases

[5] ISO/IEC WD 5339 Information technology — Artificial intelligence — Guidance for AI applications